Greenhouse Gas Observation & Climate-Smart Agriculture

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730995

SEACRIFOG Deliverable 5.1

Requirements and design considerations for an interoperable data portal

Werner L Kutsch[1], Ville Kasurinen[1], Alex Vermeulen[2], Wim Hugo[3], Lutz Merbold[4] and Johannes Beck[5].

[1] ICOS ERIC Head Office, Erik Palménin aukio 1, FI-00560 Helsinki, Finland
[2] ICOS Carbon Portal, Sölvegatan 12, SE-22362 Lund, Sweden
[3] South African Environmental Observation Network, Herzog Blvd, Foreshore, Cape Town, 8001, South Africa
[4] International Livestock Research Institute, Mazingira Centre, PO Box 30709, Nairobi 00100, Kenya
[5] Southern African Science Service Centre for Climate Change and Adaptive Land Management (SASSCAL)

Helsinki, August 2019

Outline

Executive Summary

The understanding of Earth's climate system and GHG emissions have been recognized as one of the biggest global challenges and is needed for the development of any strategy and evidence-based decision making, whether it is on food security, climate-smart agriculture, mitigation of anthropogenic greenhouse gas emissions or adaptation to climate change. Scientific analyses, reliable predictions and environmental decisions must be based on observations that can be verified and are well documented. The demand for earth observations is growing globally and therefore reliable information should be made easily accessible.

The purpose of the project "Supporting EU-African Cooperation on Research Infrastructures for Food Security and Greenhouse Gas Observations" (SEACRIFOG) is to develop a continental network of joint EU-African research infrastructures (RIs) for monitoring GHG emissions and observing climate system in Africa.

This report constitutes Deliverable 5.1 "Requirements and design considerations for an interoperable data portal" of the SEACRIFOG project. It was prepared under the lead of ICOS ERIC Head Office as a part of Work Package (WP) 5 and significant contributions have been received from South African Environmental Observation Network (SAEON), Southern African Science Service Centre for Climate Change and Adaptive Land Management (SASSCAL) and International Livestock Institute (ILRI).

Deliverable 5.1 has considered previous work carried out in Work Package 3 "Developing a common research agenda to promote Carbon, GHG and aerosol observation in Africa to fill gaps in a global observation system" and Work Package 4 "Improving technical harmonisation and data quality in environmental monitoring and experimentation. Deliverable 5.1 follows the basic philosophy of the SEACRIFOG project of designing an integrated Research Infrastructure (RI) and has focus on the e-infrastructure for the systematic data measured and provided by this RI.

Interoperable data portal storing, providing easy and fast access to reliable, high quality environmental data is a fundamental component for the future African Research Infrastructure. Deliverable 5.1 summarizes examples from existing data infrastructures in Africa and Europe representing technical and ideological frameworks and practises that can be utilized in this work. An interoperable data portal for the established Research Infrastructure Consortium in Africa would require governance structure and is not covered in this report. The future African RI should systematically contribute to global observing systems. This could be organized through the United Nations (UN) Systems (e.g. UNFCCC or UNCBD) or the Global Climate Observation System (GCOS) which is based at the World Meteorological Organisation (WMO). Last but not least there is also a need for a high-level dialogue platform (SEACRIFOG Work Package 7) which can be seen as a starting point for discussions regarding the co-operation and stakeholder dialogue between African countries and the EU.

1 Introduction

Observation, quantification and understanding of the state of the environment is a necessary requirement for the development of any strategy and for evidence-based decision making no matter whether it is about food security, climate-smart agriculture, mitigation of anthropogenic greenhouse gas emissions or adaptation to climate change. Scientific analyses based on environmental data affect decisions in all sectors of society. Reliable predictions underlying environmental decisions must therefore, be based on trustworthy, well-documented observations. Easy and fast access to reliable, high quality environmental data is therefore fundamental. The demand for earth system observation data is rapidly increasing globally, but the tools to manage, document, provide, find, access, and use such data are still underdeveloped owing to the combination of data complexity and data volumes. Each region has its specific challenges related to environmental earth observation. This study will focus on the specific requirements of a data infrastructure supporting the information flow from environmental observations to decision making in Africa.

It follows the basic philosophy of the SEACRIFOG project (see Fig. 1) of designing an integrated Research Infrastructure (RI) and has focus on the e-infrastructure for the systematic data measured and provided by this RI. This data is crucial for scientists in their quest for understanding and interpreting the complex Earth System. Environmental research data collected from in-situ and space-based Research Infrastructure contribute systematically to global observing systems often organized in the United Nations (UN) System and related to the UN Sustainable Development Goals or UN conventions e.g. on Climate Change (UNFCCC) or loss of biological diversity (UNCBD). In the case of Climate Change the Global Climate Observation System (GCOS) based at the World Meteorological Organisation (WMO) coordinates the global climate-related systematic observations including climate, atmospheric composition as well as land and ocean fluxes.

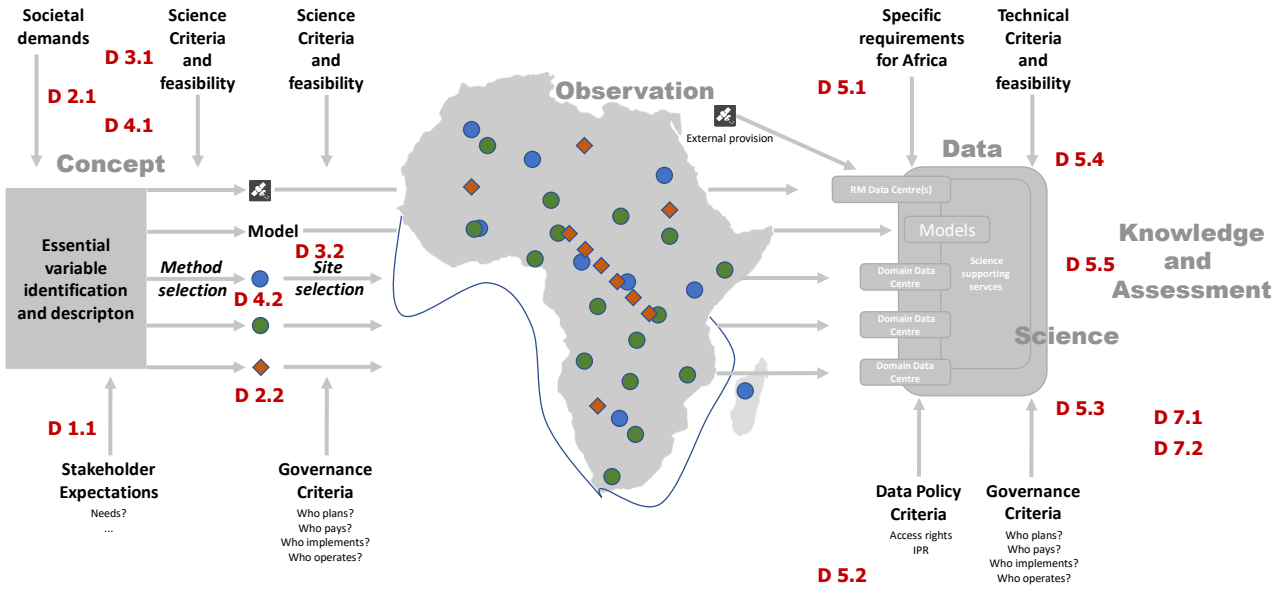


Figure 1: An overview of the general approach of the SEACRIFOG project.

It shows on the left-hand side the deliverabled that have built the design of the systematic observational system from multiple input. The right-hand side mirrors this information in the design of a related data infrastructure.

This study has been strongly built on deliverables of other work packages (particularly on the Deliverables D 3.2 and 4.2 and refer to data amounts and requirements from systematic observations estimated there. However, this study also includes strategies to make 'unsystematic data' generated in research projects or national observation programs available. There is a treasure of data in many African countries that certainly enables African scientists to support their governments and the African Union with environmental knowledge. However, it requires thorough strategies to make this treasure available.

Developing a data infrastructure underlying a systematic observation system and offering a repository for additional unsystematic data has to start with a description of requirements which is performed from different viewpoints in this study:

1. General requirements (Section 2)
2. Requirements on research data infrastructure that is unique to Africa including an analysis of the requirements defined in Del. 3.2 (Section 3, and Annex A.2);
3. Requirements resulting from Global Data Initiatives (Section 4, and Annexure A.3);
4. Existing stakeholders and potential contributors (Section 5, and Annexure A.4)
5. Implications of SEACRIFOG research into the nature and status of carbon observation in Africa, and how it can be optimally improved (Section 6, and Annexure A.5).

2 General Requirements

General requirements for data systems in support of science describe scientific enterprise in general, its governance, and general guidelines and best practice. They can be obtained from several sources:

1. **Science concerns**: questions posed by science, which, in turn, finds expression in a collection of standard variables, which if observed at the correct temporal and spatial scale, over an adequate period of time, and using sufficiently aligned protocols, can inform analysis and hypotheses; the respective scientific base of the data has been laid in the WPs 2, 3 and 4 of SEACRIFOG and their consequences will be outlined in Section 3.
2. **Governance concerns**: beyond stable funding and location, this is largely finding expression in open data considerations, by extension requiring all tax-funded research output to be openly and freely available unless some specific limitations apply, resulting in policies, and best practice guidelines, which are briefly mentioned here and more broadly explored in D 5.2. In addition, Open Science promotes the reproducibility and veracity of science. Includes sustainability, funding, and trust concerns, and may influence decisions on topology - for example whether systems should be federated or centralised.
3. **Informatics concerns**: finding expression in considerations such as reference models, information models, vocabularies, standards and specifications, and architectures.

More specific challenges often arise from a mixture of these concerns. Tab.1 gives an overview of sources that have been used to describe the major requirements of a data infrastructure in the context of SEACRIFOG-designed observations.

#	Nature of Concern	Description	Source
1	Informatics, Science	Common characteristics of environmental research infrastructures in Europe, and a reference model derived from these - developed by the European Environmental Research Infrastructures during the Projects ENVRI and ENVRIplus[1].	ENVRI Reference Model
2	Informatics, Science	Principles of accessibility, interoperability, re-usability, and discoverability.	FAIR Principles as introduced by the FORCE 11 group
3	Governance, Informatics	Open Data and Open Science considerations, as well as recommendations from bodies such as RDA. To date, RDA have published a limited set of adopted recommendations, but several ongoing initiatives are of interest.	RDA Recommendations Open Data Open Science RDA Working Groups
4	Science	The scope of variables to be observed by a Carbon Observing Research Infrastructure in Africa, as well as the protocols, data families, current and optimal future observation locations, and platforms associated with these.	SEACRIFOG Work Package 4: Key Variables Data Requirements Protocols

2.1 The ENVRI Reference Model

ENVRI Reference Model has been developed based on analysis of representative environmental research infrastructures in Europe (Chen 2013). Based on their computational characteristics five common subsystems have been identified and the fundamental basis of this division is due to observation that applications, services and software tools are designed and implemented around five major physical resources. These resources are the sensor network, the storage, the (internet)

[1] We include the implications of ENVRI in this section - even though the focus was on environmental research infrastructures, the reference model is generalised in nature and applies to most data infrastructures.

communication network, application servers and client devices. The definitions of the five subsystems according to Chen (2013) are described in ICOS data life cycle description as follows:

- Data acquisition: collect raw data from sensor arrays, various instruments or human observers, and brings the measurements (data streams) into the system
- Data curation: facilitates quality control and preservation of scientific data. It is typically operated at a data centre
- Data publishing: enable discovery and retrieval of data housed in data resources managed by a data curation subsystem
- Data processing: aggregates the data from various resources and provided computational capabilities and capacities for conducting data analysis and scientific experiments
- Data use: supports users in gaining access to data and facilitating the preservation of derived data products

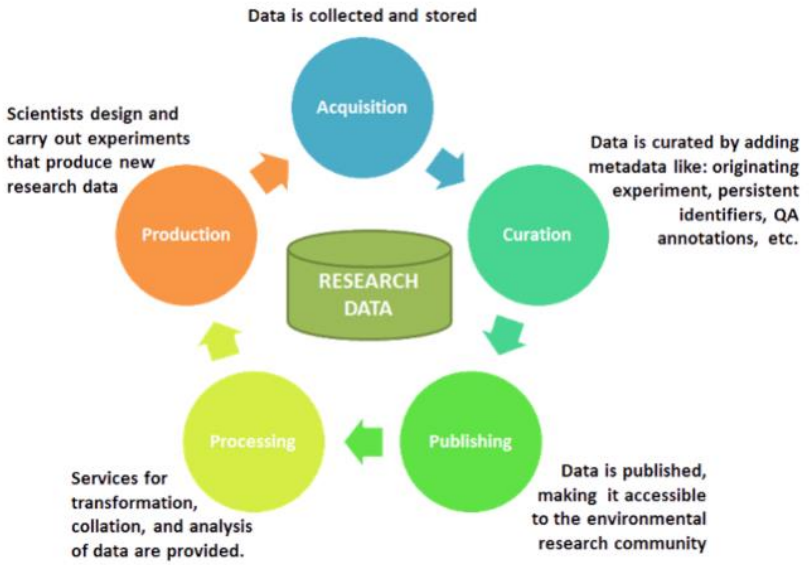


Figure 2: The 5 phases in the data life cycle according to the ENVRI Reference Model v2.0.

The ENVRI Data Life Cycle has implications for systems that have to support it: these are summarised in Annexure C.

2.2 Open Access and FAIR Principles

The past decade in research landscape has shown that the importance of open access for data has been increasing internationally (Hugo, 2019). Stakeholders and funding agencies require that

publicly funded research programs have to share the data collected during research projects and the amount of studies utilizing data repositories or data infrastructures is continuously increasing. The FAIR principles – introduced in 2016 by the Force 11 Group – can be seen as a concept to ensure that as many users as possible can benefit from the investments that have been allocated to satellites, sensors and environmental monitoring strategies. The FAIR principles elaborate more than open data access to data for the scientific community and researchers. They comprise four attributes open data should have: Findable, Accessible, Interoperable and Re-usable. These attributes are described in more detail in the following subsections.

2.2.1 Findable: Metadata

The elements of FAIR principles are related, but independent and separable. The first requirement for FAIR is that researchers or stakeholders interested in specific data must be able to find it (Discoverability). Data can be made available without established infrastructures. Starting point can be, in its simplicity, that data object or data set and metadata describing data can be found through web page (URL) or its location is described in service catalogue listing available data sources related to specific themes. Data objects are findable when metadata attached to each data object exists. If metadata is provided in machine readable format it can be searched through search engines and service catalogues.

In established data infrastructures that has agreed and defined principles regarding data acquisition, curation, publishing, processing and use can create processes that generates Digital Object Identifiers (DOI) or Persistent Identifier (PID) for data objects. Using these practises infrastructure can confirm that its data objects are cited in the same way. However, DOI and PID generation is not mandatory for findability of data and in some cases data ownership may limit the ability of infrastructure to sign persistent identifiers. In SEACRIFOG project Collaborative Inventory Tool has fulfilled mainly the requirements related to findable data objects by providing a catalogue of available data products and sets for African context. However, improvements are needed for example related to metadata import and export methods (Deliverable 4.2).

As indicated in SEACRIFOG deliverables 4.1, 4.2 and expressed in D4.3 recommendations the future African RI should have harmonized structure for metadata where the minimum requirements for essential variables are agreed. To achieve this requirement, local measurement networks and contributors will need support for capacity-building to understand and implement agreed guidelines. In current situation quality of metadata might cause issues for data portal implementation and detailed analysis regarding similarities and differences between SASSCAL, SAEON, SEACRIFOG and other data providers will be needed.

Metadata requirements are in some degree dependent on specific measurements and used techniques. SEACRIFOG deliverable 4.3. has identified and listed in total 140 metadata protocols that can be utilized when designing African RI data portal. Deliverable 4.3. assessed 82 protocols that deal with ground-based and sea-borne observations. Metadata requirements are also dependent on the measurement platform and requirements may vary between different measurement techniques.

2.2.2 Accessible

When data object is findable and metadata describing the object exists, data cannot be used without access to it. Data can be findable but not accessible or accessible but not findable. Depending on intellectual property rights and used data licenses access can be granted after registration or accepting terms and conditions. Data can be available upon request and in some cases applied licences may restrict to its use to for example non-commercial purposes.

Data portals or data infrastructure can maximize the benefits that can be reached through the use of its data by providing open access to users willing to utilize available data resources. For example, Creative Commons Attribution 4.0 international licence (CC BY 4.0) allow user to share, copy, adapt, transform and redistribute the material if user gives appropriate credit, provide a link to licence and indicate if modifications to the original data were made. ICOS Carbon Portal is following CC BY 4.0 licence and used technical solutions are based on open source principles.

Typically established research infrastructures has common data policy, which defines principles regarding registration, accessibility and data use. In data infrastructure Findability (F; metadata) and Accessibility (A;) are the first two conditions that are required for interoperability. When these previous two requirements are fulfilled in design considerations of a data portal, it will create a good grounding for interoperability and reuse of data objects. Both machines and humans should be able to judge the actual accessibility of data objects.

2.2.3 Interoperable

If data portal is designed to consider requirements of Interoperability, it will enable accessing and processing data objects from multiple resources. To confirm that technical requirements related to interoperability of a data portal, it is essential to apply international recognized standards and support widely used data formats (csv, JSON, netCDF, OGC Services, SensorWeb and SensorThings, etc). If technical requirements are reached, data can be then used for mapping, visualization and other representations and analysis. Metadata and data objects must be machine readable format and available from machine to machine interactions that enables people to find, explore and understand the structure and content of data objects.

2.2.4 Re-usable

In some disciplines data life cycle is considered to be very short if data is collected for one specific case-study, which cannot be repeated. However, this is rarely the case with data products related to environmental monitoring. The recognized Essential Climate Variables in the SEACRIFOG project creates a fundamental basis for environmental monitoring and science-based decision making.

Re-usability of data can be achieved when data objects are compliant with three first FAIR-principles. Metadata describing data objects should be rich and well-described so that in can be automatically or with minimal effort linked and integrated with similar other data sources. Re-usability of analysis means that used data and utilized resources such as code and algorithms should

be linked data objects and allow others to reproduce specific work. Reproduction creates therefore a basis for continuous development of analytical frameworks and supports the development of open science landscape. In established and operational data infrastructures published data objects can refer to their sources and enable proper citation to used data, metadata and used methods.

2.3 Conclusion on the FAIR principles for the project

A large amount of important work has been carried out during the SEACRIFOG project, which supports directly open access and FAIR principles. Especially deliverable 4.2 Climate Change Observations across Africa: "Data Requirement and Availability" has succeeded to summarize the current situation related to spatial and temporal data coverage and best practices in Africa. SEACRIFOG WP4 has identified Essential Climate Variables needed to develop science-based strategies to improve food and nutrition security and climate change mitigation. Furthermore, the SEACFRIFOG Collaborative Inventory Tool (seacrifog-tool.sasscal.org) serves to systematically capture information on relevant variables, observational infrastructures, existing data products and methodological protocols that are linked to greenhouse gas emissions and food security across the African content and surrounding oceans.

The recent progress in SEACRIFOG project regarding existing datasets and -products provides a good basis for open access and FAIR principles implementation in Africa. The SEACRIFOG consortium partners has for asked and received contribution from local and international measurement networks. At the time of writing the SEACRIFOG Collaborative Inventory Tool contained metadata on 142 datasets or -products. Together with SAEON and SASSCAL data portals, SEACRIFOG projects supports open access and FAIR principles implementation related to Essential Climate Variables (ECV) in African continent scale, which is needed for science-based decision making.

Altogether, the future African RI needs to define structures and practices that follow FAIR principles and ensure interoperability with European and Global research infrastructures.

3. Project-specific and Africa-specific considerations

Scholes et al. (2019; Deliverable 3.2 of SEACRIFOG) provide a detailed lists of data products related to the variables derived by Beck et al. (2018; Deliverable 4.1 of SEACRIFOG and 2019; Deliverable 4.1 of SEACRIFOG). They have been binned to several approaches that constraint the design of a respective e-infrastructure that should be developed in close connection to the observational system.

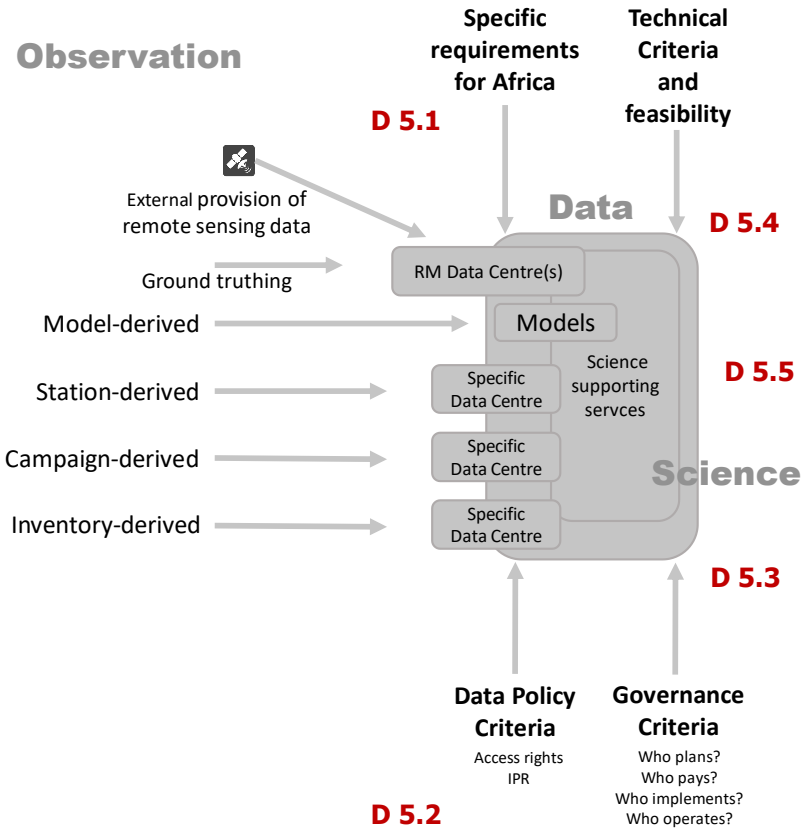


Figure 3: A closer view on the required elements of an e-infrastructure related to an observational system on GHG in Africa.

It should comprise different elements that could be combined in one location or be distributed and virtually connected. Those are:

- A centre for remote sensing data storage, their basic processing and for the derivation of data products on defined variables (i.e. RCMRD in Kenya).
- A modelling centre running and further developing models for data integration and estimation of variables that are too complex, laborious or expensive to measure while also allowing for data mining approaches.
- Specific data centres for archiving, QC and processing of in situ data from stations, campaigns and/or inventories.
- The common user interface should ideally be combined with a number of web-based tools for data analysis in order to support scientific users as well as to include the wider public (schools, NGOs, individuals) and developers that could create services based on environmental monitoring data.

The future African RI responsible for facilitating and maintaining a data infrastructure as outlined here that receives, archives and provides data measured by the RI to users, and provides services to scientists will need clear policies, formal administration, long-term funding and thorough design and location decisions. It will require co-operation at African continental scale and high-level dialogue between participating countries who volunteer to facilitate the operations of the RI. One possible platform to start such work could be co-operation between the African and the European Unions, with support by WMO and building on existing RIs such as the Integrated Carbon Observation System (ICOS) in Europe and SAEON/SARIR EFTEON in South Africa and service centres such as WASCAL and SASSCAL. The technology, knowledge and experiences from these partners are outlined in Section 5 of this study. A blueprint for a data infrastructure including different

organisational options and estimates on necessary computing and storage capacities, personnel and costs will be described in more detail in SEACRIFOG Deliverable 5.4.

Last but not least, some aspects of data access and ownership bear some Africa-specific requirements. Many variables identified as essential for the envisaged observational research infrastructure require expensive high technology instrumentation and the respective scientific knowledge. Collaboration with researchers and research institutions from across the globe or with commercial service providers will, therefore, be highly probable at least in the implementation phase and the first years of operations. Respective cooperation agreements need to guarantee full access to all data by countries hosting the measurements, ideally through joint ownership, meaning that each of the joint owners shall be entitled to use their jointly generated and jointly owned data and research results, whether patentable or not, for non-commercial research and teaching activities on a royalty-free basis, and without requiring the prior consent of the other joint owner(s).

3.1 Previous experience from Africa-specific projects

During 2013/14, SAEON conducted work on behalf of ICSU-WDS in respect of establishment of a 'Network Data Centre' in Africa. The main aim of such a Data Centre would be to serve as a focus point for a federated data infrastructure in Africa, and to provide data repository and curation services to such participants that could not develop their own contributing infrastructures. A workshop was conducted during the 5th African Conference for Digital Scholarship & Curation, 26-28 June 2013 at the University of KwaZulu-Natal (UKZN), Durban, and results from this workshop were used to identify the major impediments to the establishment of such a centre.

The main findings were:

1. Establishment of a Network Data Centre will require participation of a Trusted Digital Repository or Data Centre as a minimum infrastructure component. South African Research Data Infrastructure is likely to be adequately provisioned to cope with additional deposits from, for example, the SADEC region.

2. A variety of policies, access models (including Open Access), data quality, and data preservation combinations are likely, and rather than to limit the scope of referenced data sets by making the bar too high (and aligned with developed world norms), we are proposing an incremental approach to accommodate the variety, migrate to more mature or ideal levels of performance, and promote inclusion. It is, in our opinion, better to reference data irrespective of quality, license, or preservation status than to render it inaccessible. Filters can then be applied to expose compliant data sets to ICSU-WDS harvesters.

3. In the absence of member policies, the ICSU-WDS Data policy[2] will be proposed, together with an Open License (Creative Commons-based)[3]. The Open Licenses will have to be

[2] http://www.icsu-wds.org/services/data-policy
[3] http://creativecommons.org/licenses/by-sa/4.0/

supplemented by a small number of standard licenses dealing with valid restrictions (ethics and privacy, commercial data, and government classified data).

4. Funding is (and is likely to remain) a major impediment. We are proposing a way forward that makes maximum use of voluntary contributions, both of expertise and infrastructure, and ask that ICSU and WDS give serious consideration to the development of a framework within which such contributions can be acknowledged, managed, and structured.

Not identified in the survey, but nevertheless pertinent, are concerns in respect of data sovereignty.

The respondents interviewed within a stakeholder group as part of SEACRIFOG Work Package 4 activities also identified some important technical impediments:

#	Concern or Issue	Discussion	Reference
1	High Costs	Both for implementation and maintenance	Ballesteros (2018)
2	Energy Availability	Energy costs are often high, and may not be easily available in locations where observation infrastructures are located – complicating data transfer and primary backup.	Ballesteros (2018)
3	Connectivity	Connectivity is a problem - not only in rural areas, but also in some urban situations. Costs are high.	Ballesteros (2018)
4	Human Capacity	Specialised companies supporting instrumentation and IT infrastructure is a constraint, as well as suitably qualified technical and scientific personnel.	Ballesteros (2018)

The study also identified desirable characteristics of the data products and services that such a RI should offer, the following

#	Concern or Issue	Discussion	Reference
5	Adaptation Focus	In the developing world, focus in terms of responses to climate change favours adaptation rather than mitigation. This is, in part, because the developing world is at the moment a lesser contributor to global GHG emissions.	Ballesteros (2018)
6	Mediation	Translation of evidence and scientific findings into implementable policy, planning, and decision support.	Ballesteros (2018)

4 Requirements from Global Initiatives

4.1 Landscape Review

International trends, drivers, and guidelines determine a large number of design considerations in respect of e-infrastructure. These include the following, and are based on several references:

1. **Architecture-related:**

    a. **Standards and Specifications for Metadata and Data Services**: these depend on the scope of data families (Mirtl et al. 2018) required by SEACRIFOG in respect of operational data management, as well as those implied by agreed exchange mechanisms (See Annexure A);
    b. **Data Management and Curation**: guidance from FAIR, GEO, WDS, RDA, CoreTrustSeal, and others influence some of the elements of infrastructure design, and specifically impact on workflow arrangements, the curation state of research outputs, and elements of quality assurance and trust;
    c. **Discovery and Search**: services, user interfaces, and broker integration for dissemination of data within multiple communities;
    d. **Reliable Citation**: required for proper attribution, determination of dependencies, and increasingly for provenance, as well as linkages to scholarly publication workflows - refer to DataCite and Scholix for examples;
    e. **Application**: access to data via standardised services, subsetting and query facilities, inclusion into scientific workflows, and data visualisation and exploration tools;
    f. **Rating and Metrics**: frequency of use and feedback on data and metadata quality, user annotations and crowdsourced data, and quality assurance of crowdsourced contributions.
    g. **Semantic Web, Controlled Vocabulary, and use of Persistent Identifiers**: Identifiers need to be implemented as a minimum for the following aspects of research data and its management:
        i. **Samples and specimens**, including digital samples;
        ii. **Research output** (scholarly publications, data - including dynamic data, code and algorithms, protocols and methods) (see RDA recommendations - Rauber et al., 2016);
        iii. **Researchers**, **institutions**, **repositories**, **funders**, and **projects**;
        iv. **Instruments and sensors**, including virtual instruments, which has an overlap with **protocols**.
        v. **Vocabularies**, **ontologies**, and **thesauri**, implemented as standardised **name services**, are required to describe
            1. **Semantic**, **temporal**, and **spatial** coverages - of which semantic coverages and can benefit from the development of standard variable collections, such as already in process for climate, ocean observation, and biodiversity, *and more specifically, linking these variables explicitly to one or more* ***societal benefit areas***;
            2. **Processes**, characteristics of data sources and data sets, and similar qualitative information - for example as developed as for biological collections (Humboldt Core);

3. The structure of SEACRIFOG and its **sites**, including information on the scientific context of the site, length and scope of observation, and hopefully aligned with information currently collected in DEIMS-SDR as used for ILTER.

2. **Practical Guidance and Systems Engineering Considerations:**

    a. **Granularity of Metadata** and **Citation** for **Dynamically Updated Datasets**: guidelines are available for the management of persistent identifiers associated with dynamically updated datasets. These are of critical importance to SEACRIFOG, given the continuous nature of many of its observation sites;
    b. **Modularity** of Loosely Coupled Services and Interfaces: There is significant support for a service-oriented architecture for global research data infrastructure - this allows a 'plug-and-play' based composition of applications and portals, utilising contributions from many;
    c. Interoperability
        i. **Syntactic**, **Schematic**, and **Semantic** Interoperability standards need to be implemented, as discussed in Annexure B;
        ii. **Brokers and Mediators**: In practice, standardisation is never perfect, and mediation of any of the above is usually required. SEACRIFOG is likely to require a brokering service to aggregate metadata from all participants and to allow data access in a federated system of systems.

3. **Open Data and Open Science**: Considerations include publication of data in support of scholarly outputs, minimum metadata requirements, licenses, data citation and citation indices, and integration with scholarly publication workflows.

4. **Reproducibility and Trust**: Formal certification of trusted data repositories is already available and serves as a benchmark for sustainable infrastructure, with indications that such certification is likely to be extended to other aspects of scientific output (samples, code, vocabularies, and protocols).

The implications of these considerations for SEACRIFOG is summarised in Annexure A.3

4.2 Generic Use Cases

Figure 4.2.1 provides a summary of the generalised Research Output (Data/ Code) Infrastructure (ROI) use case. This so-called 'Publish-Find-Bind' architecture or use case is central to most of the infrastructure that SAEON develops and maintains.

- Firstly, it is assumed that research outputs (**Data**, digital objects, **Code**, etc.) will be *published* with **Metadata** that is sufficiently complete to allow proper citation, discovery, and re-use. The effectiveness of the step is largely a function of the quality of **Curation** that is performed (data quality assurance, completeness of metadata, proper preservation, etc.) ("*Publish*").
- For this to work well, both data and metadata needs to be standardised.

- Applications, systems, and end-users can then **discover** ("*Find*") data via any number of standardised search protocols (for example, CS/W, OpenSearch, and OAI-PMH). This discovery process may lead citation in its own right, and may in turn be influenced by citation indices and community ratings or evaluations of the research outputs.
- Once research outputs have been found, they may be put to use ("
- *Bind*"): generally, these will be one of three actions: **accessing, downloading, or streaming** the data (in which the concept of **content negotiation** could play a role), invoking **code** or **processing/ transforming** data via services, or alternatively **visualising or exploring** data. Such action will also require citation in most cases.
- Binding research objects to useful application often requires Semantic
- **Annotation** - in essence, tailoring the input to fit some form of real-world application. Such annotation could be simple (selecting and naming the variables to be used in a chart) or very elaborate (mapping a number of data sources and transformation processes to support a specific decision support tool).
- Finally, it is unusual for all elements of the architecture (data and data services, code and processing or transformation services, discovery services, and metadata documents) to all comply fully with specifications, and in some cases it may be required to
- **mediate** this heterogeneity. Doing so in a managed or optimised manner is referred to as **brokering**.

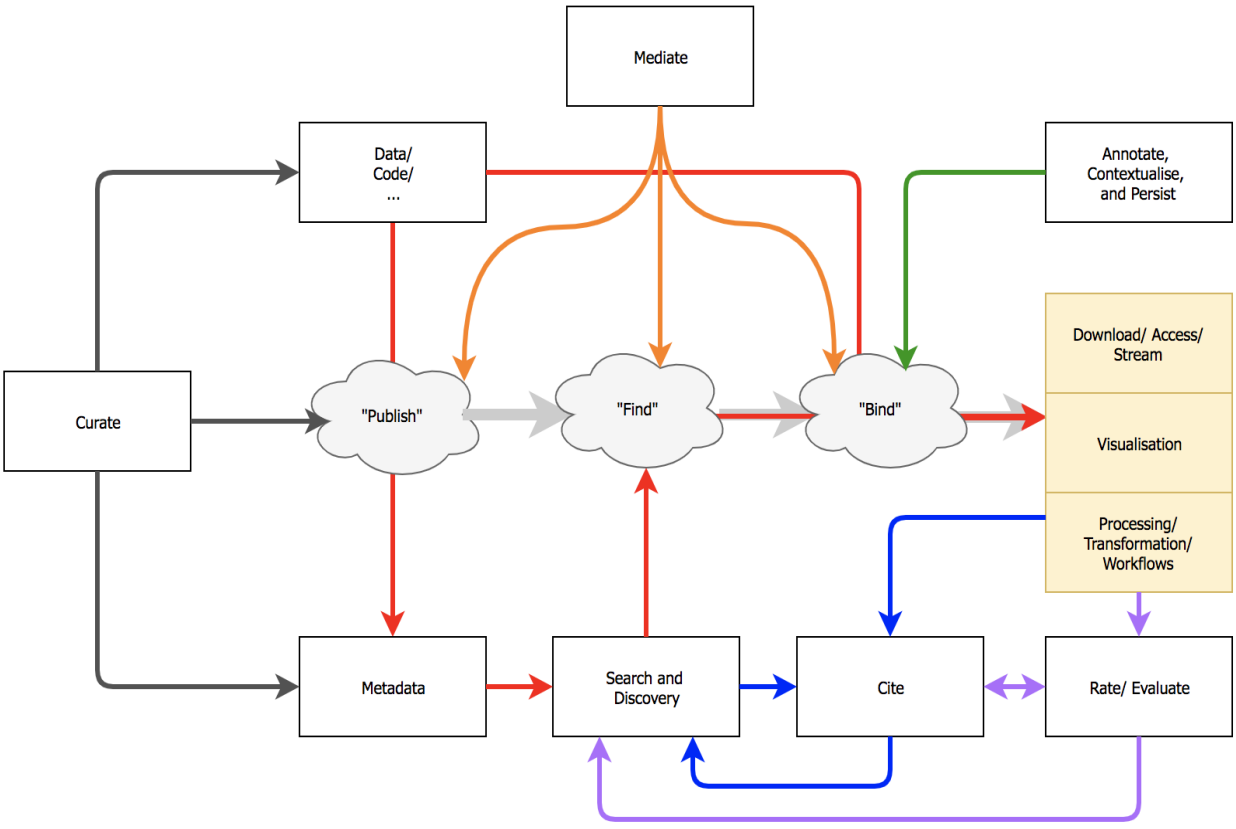


Figure 4.2.1. Generalised Research Data Infrastructure Use Case.

Not all the systems deal with research output: there is also static or semi-dynamic contextual information (such as associated with the general pages and content in a website), business objects such as reports, personnel profiles, contracts, and similar, and links to vocabularies or name services that are used to formalise the semantic web (
Linked Open Data).

Based on the above, and taking note of relevant literature defining the role of e-infrastructure in environmental observation networks, we can summarise functionality for the following use cases (Fiore et al. 2015, Mirtl et al. 2018, Lopez-Ballesteros et al., 2018, Hardisty et al. 2019):

Use Case	**Description**
Registration, authentication and identification of end-users	While not a requirement for access to services and data provided under open licenses, it is nevertheless useful to keep track of users for purposes of motivation and measurement of utility, enhancing user experience, and managing access to content. Implementation can be based on open identification systems such as OpenID and EduGain.
Administrative Functions	Allowing end users to register and manage definitions of participating networks and their observation infrastructure (as currently implemented in DEIMS-SDR), linking network or institutional metadata collections for synchronisation, and managing user-generated objects and content.
Search and Discovery	Allowing end users to search for networks, infrastructure, instruments, data objects and services, and to persist predefined search definitions for future use. High-end faceted search capabilities such as ElasticSearch or SOLR can be employed in this role.
Application	This includes binding of data services to user applications and processes, data exploration and visualisation, inclusion of data services into distributed web processes, and composition of objects based on distributed services (for example Atlases or time series visualisations based on multiple, distributed services), and linking of standardised data services into VRE workbenches and notebooks.
Publication	The act of publication in a distributed environment needs to be managed, and publication workflows are potentially required to assist with this.
Curation	Users that contribute content need to be in a position to manage such content through curation workflows - forming a large part of the technical aspects of certification as trusted infrastructure.
Assessment and Rating	End users need to provide feedback on quality of metadata and data, and information on formally gathered metrics is required (AltMetrics, Scopus, Web of Science, and others).

Citation	Allow end users to reliably cite data and other digital objects obtained from the infrastructure (for example via DataCIte), even if the data was subsetted from a large or dynamic dataset, or was compiled from multiple data sources distributed in the web.
Semantic Linking and Annotation	Users require secondary tools for contextual annotation of data, linking keyword and other descriptions to globally recognised vocabularies, and enhancing search operations.

5 Existing Data Infrastructure

5.1 The SAEON Data Portal

The SAEON Data Portal was first developed in 2008 to serve as a shared data and metadata repository for both SAEON and the CSIR. Since then, significant development has followed in support of a wide variety of stakeholder initiatives and shorter-term projects, all contributing to establishment of a service-driven, standards-based infrastructure that continues to grow. We now refer to this infrastructure as the SAEON Open Data Platform (ODP), and it provides not only hardware and software infrastructure, but also soft infrastructure such as guidance, best practice, curation, and support services.

The platform is used for publication, discovery, dissemination, and preservation of Earth and Environmental Data, chiefly with funding from NRF, Department of Environmental Affairs, and Department of Science and Technology. This platform hosts several portals[4] and gateways[5], including SARVA, The South African Earth Observation System of Systems (SAEOSS), the BioEnergy Atlas, and SAEON's own data portal. It also serves as a platform for hosting the South African Spatial Data Infrastructure (SASDI), and has been used for internationally funded exploratory work to establish Africa-wide prototypes for data management in the domains of biodiversity, human health, and socio-economic sciences. As a result, a large ecosystem of systems, applications, and services have emerged, utilising a shared metadata catalogue and components.

The ODP operates on a principle of mutual benefit, and by design is capable of providing access to metadata and data or digital content across all portals and gateways, as well as allowing improvements and extensions funded by a specific initiative to be available to other potential users at low or no cost, depending on their requirements.

[4] Portal: providing theme or community-specific access to resources

[5] Gateway: providing generalised user, service, and system interfaces to resources

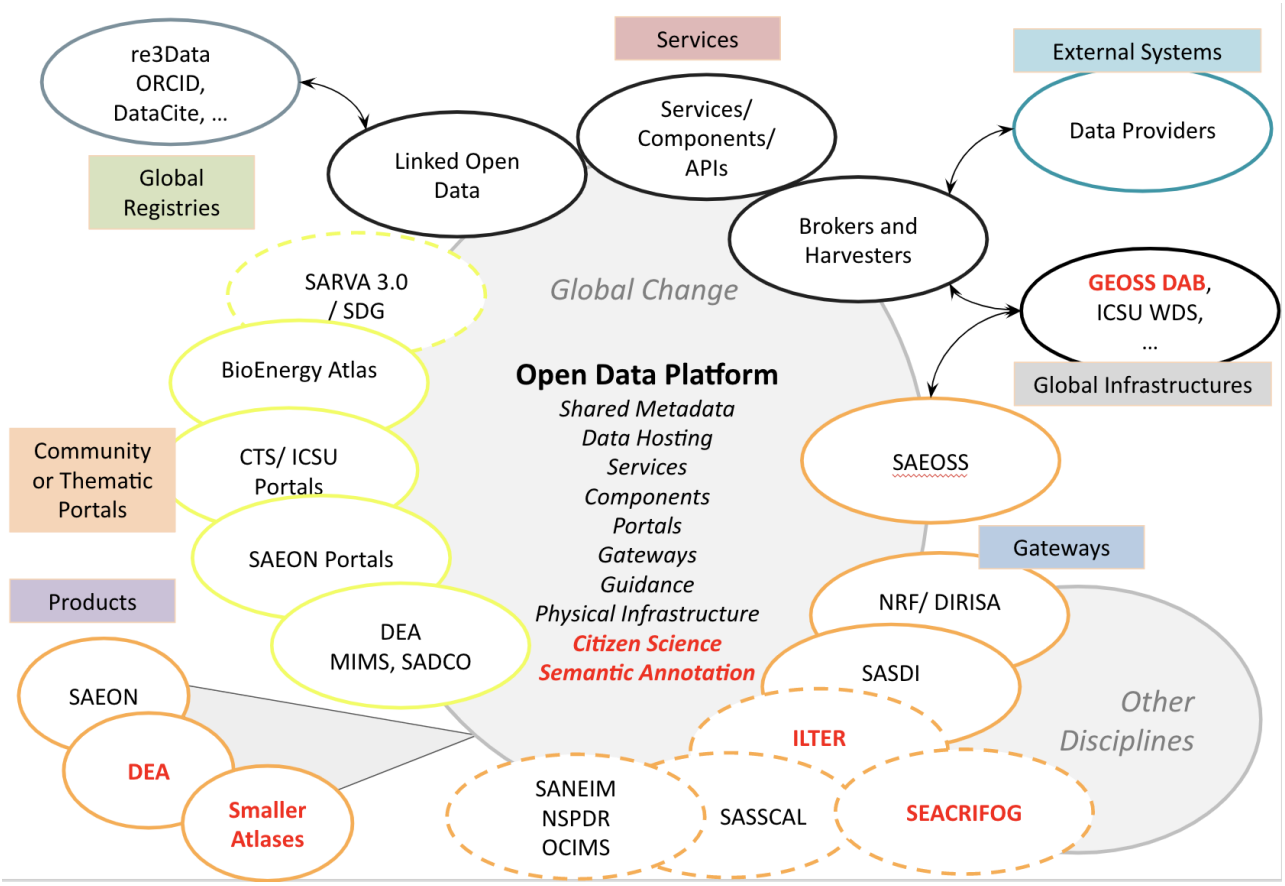


Figure 5.1.1: National and Global Research Data Infrastructure Context
Key: GEOSS DAB - GEOSS Data Access Broker, ICSU WDS - World Data System, SAEOSS - South African Earth Observation System of Systems, DIRISA - Data-Intensive Research Infrastructure for SA, ILTER - International Long-Term Ecological Research, SANEIM - South African Environmental Information Metadata System, NSPDR - National Spatial Planning Data Repository, SASDI - South African Spatial Data Infrastructure, CTS - CoreTrustSeal, SARVA - South African Risk and Vulnerability Atlas, SDG - Sustainable Development Goals. Smaller Atlases: Terrestrial Carbon Sinks Atlas, Agro-hydrological Atlas, SAWS Climate Atlas.

The ODP allows any number of ***harvesters*** (capable of brokering several mainstream metadata standards and service protocols) to be configured for any portal that it supports, and as such can automatically synchronise metadata collections from as many contributors as needed. With the operationalization of SASDI, this portfolio will grow to include most government departments.[6] Over time, several research institutions and contributors have been added to the portfolio as and when required. Our experience, though, has been that automated harvesting is the exception rather than the rule, and we have subsequently implemented several additional mechanisms to assist stakeholders with metadata exchange.

SAEON now operates significant physical infrastructure in its own right (up to 300 TB of online storage, split between operational, test, and failover/ disaster recovery facilities), and the ODP allows rapid deployment of new portals and gateways at relatively low cost. This provision will grow to roughly 2 PB during 2019/20 in anticipation of a large volume of climate modelling, video

[6] SASDI Act, http://www.sasdi.gov.za/About/SDI%20Act.aspx

observation, and time series data from new investments in instrumentation and modelling capability.

SAEOSS serves as a gateway to GEOSS (GEO[7] System of Systems) through the GEOSS Data Access Broker, exposing locally produced research outputs to a global user base, and in principle affording South African researchers access to globally available data sets.

The components for exposing specific quality assured data sets to the ICSU World Data System[8] (WDS) are also in place, and once other aspects of sustainability and governance have been addressed, accreditation of selected portals within the ODP will be sought via CoreTrustSeal. This accreditation serves as recognition by peers that the data platform is properly managed, serves quality assured data, and will be available for the foreseeable future.

Finally, the technical and licensing aspects of issuing data sets with Digital Object Identifiers[9] (DOIs) via DataCite have also been addressed. This allows data sets to be published internationally and for data sets to be cited reliably in scholarly publications. Datasets thus become formal scientific outputs that attract a citation index.

5.1.2 Stakeholders

SAEON serves many external stakeholders in addition to the programmes managed internally - these being SAEON's own science programme, the Shallow Marine and Coastal Research Infrastructure (SMCRI), and the Extended Freshwater and Terrestrial Observation Network (EFTEON). We collaborate directly with the Department of Environmental Affairs to provide ICT and data infrastructure services to the Oceans and Coasts Research Directorate via MIMS, SADCO, and, in future, OCIMS. We also assist the Department of Environmental Affairs with development of an integrated Climate Change Information System, aligning and linking a portfolio of disparate reporting, monitoring, and evaluation systems for our National Climate Change Response. SAEON has been hosting the South African Spatial Data Infrastructure (equivalent of INSPIRE in Europe) since developing it for the Department of Rural Development in 2015. We serve a number of local stakeholders and collaborators in various ways, including universities, research councils and agencies,

5.1.3 Architecture

SAEON has now implemented a multi-layer architecture that is strongly aligned with internationally and community adopted standards and specifications for both data and metadata services.

The architecture is built in the following distinct layers:

1. **Data Interfaces**: increasingly automated data sources, of which the remotely sensed stream (Sentinel Hub) and instrument time series are currently the most mature. Work has started on citizen science and crowdsourcing platforms integrated with iNaturalist, and on social

[7] Group on Earth Observations - http://www.earthobservations.org/index.php
[8] http://www.icsu-wds.org/
[9] http://datacite.org/

media and news item scraping services. Massive growth is underway in automated platform observation: multispectral, LIDAR, and visible spectrum data from drones, aircraft, camera traps, and underwater vehicles.
    - Interfacing this layer to data stores is a major future focus - automated methods for quality assurance, calibration or transformation, infilling, feature recognition, and clustering being prominent.
2. **Data Stores**: SAEON is envisaging a number of distinct data stores, and is aligning our work with the community to develop the concept of data families (Mirtl et. al, 2018). The most mature of these are spatial data and time series data stores, with work underway for file/ object and multidimensional data stores. Significant effort will go into development of media and point cloud management based on automated observation platforms.
    - A major challenge to be addressed involved development of automated metadata creation and synchronisation for continuously updated data stores. For this, we hope to implement PID[10] management in line with RDA recommendations (Rauber et al., 2016)
3. **Metadata Management**: SAEON is ready to convert from the current release of our metadata management layer (based on Plone and PostGreSQL) to a newly developed infrastructure based on CKAN and ElasticSearch. Both the existing and new releases are integrated with DataCite for DOI registration. We also need to extend our metadata catalogue and citation capabilities to archive materials and corporate digital object management - but this is less important for our scientific applications. Harvesters need to obtain metadata from a variety of source schema, and internally, we convert these to an extended DataCite Schema.
    - Metadata is available for harvesting and search queries via standardised APIs (OAI-PMH, and CS/W). All metadata records, wherever possible, include references to standardised data services. These vary by data family, and are reflected in detail in Annexure B. We have also started implementing ODATA and GraphQL REST APIs for our bespoke Relational Database Systems and vocabulary services.

[10] Persistent Identifiers

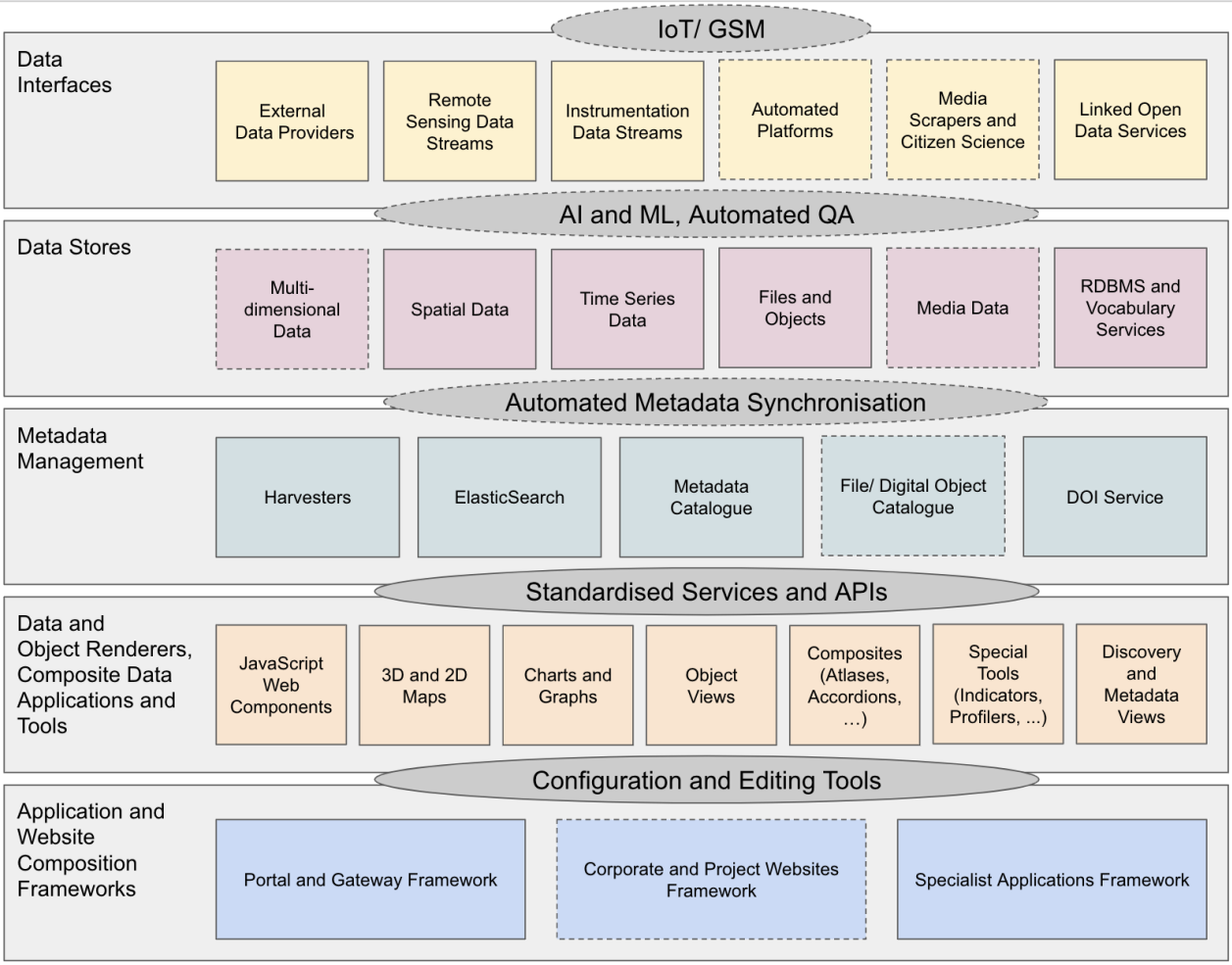


Figure 5.1.2: SAEON's Layered Systems Architecture. Dotted Lines - In Process

4. **Generic Renderers**: SAEON build applications by preference using a portfolio of generic data and object renderers, and these are, in turn built using standard web components (based on Bootstrap, React, Ant.Design, OpenLayers, WorldWind, and D3). These components are capable of visualising configured and semantically annotated standard data and metadata services, producing charts, graph (network), map, object, discovery, and composite views. Furthermore, JSON-based configurations are used to create composite views (for example Atlases as a composite of individual maps, linked to standardised data services). More elaborate tools are also built, allowing visualisation and exploration of more complex visualisation aggregations such as multicriteria indices, profilers, and indicator views.
    - A major thrust, currently underway, is aimed at building user-friendly composition tools that can assist with configuration of the renderer objects.
5. **Application and Website Frameworks**: Once again using mostly JSON-based configuration, we utilise three different standard frameworks for website and web application creation. (1) A lightweight, configuration driven framework for deployment of portals and gateways, using largely JavaScript and React, (2) a WordPress and Javascript-based framework for corporate, node, and project websites that require significant contributions from non-technical staff, but can include data-driven components and renderers, and (3) a framework for deployment of bespoke web applications with a significant RDBMS or community specific specialisation.

5.1.4 Outlook

SAEON recently cemented its funding regime for research data infrastructure and as a result, we are now in a position to update and operationalise many of our services and applications that are outdated or require rework to modernise and publish the underlying code. This process is expected to continue for another 12-18 months. IN the interim, we are ready to re-launch our SAEON Data Portal.

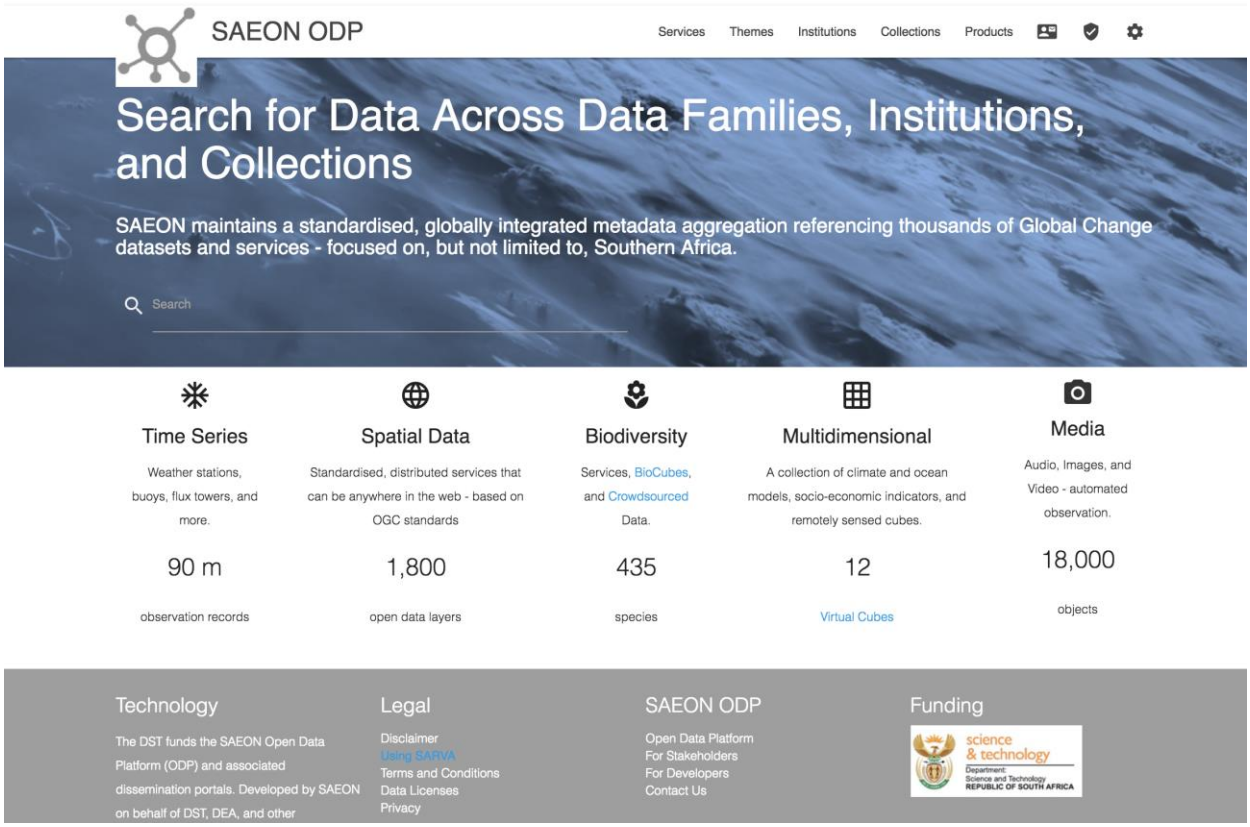


Figure 5.1.3: Beta Version of New Release for the SAEON Data Portal - In Process

SAEON is also in process of establishing a gateway that will consolidate all metadata that meets the following criteria, in addition to being FAIR:

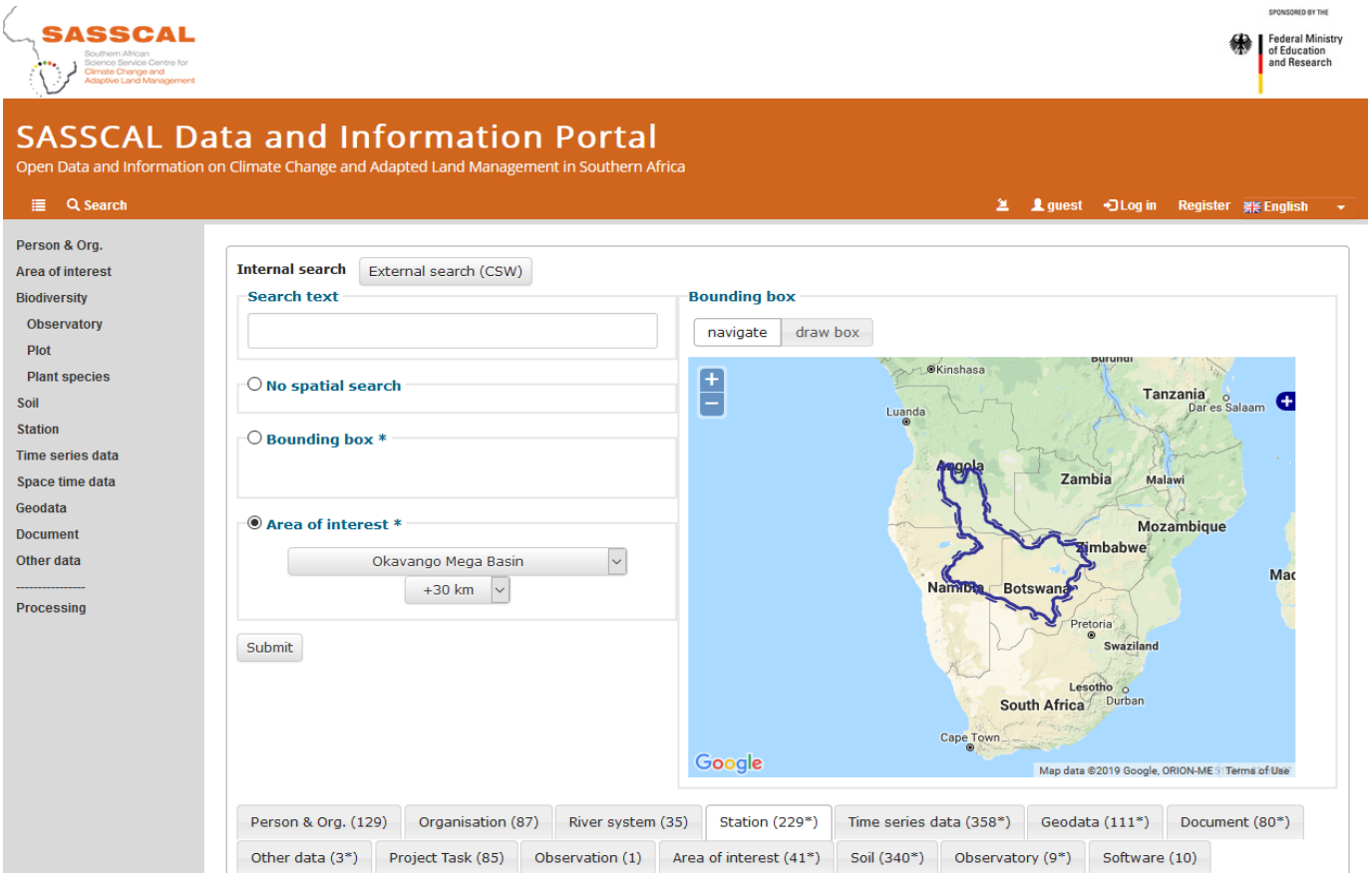
1. Open licenses - preferably Creative Commons 4.0 BY-SA;
2. Standardised data services;
3. DataCite DOI available;
4. Quality assured data;
5. Related to Global Change

This subset of our collection will be presented as a South African Global Change Data Centre, and CoreTrustSeal certification as a trusted repository is being sought.

5.2 The SASSCAL Data and Information Portal

The Southern African Science Service Centre for Climate Change and Adaptive Land Management (SASSCAL) is a joint initiative of Angola, Botswana, Namibia, South Africa, Zambia, and Germany in

response to the challenges of global change. SASSCAL aims to strengthen the Southern African regional capacity to generate and use scientific knowledge products and services for decision



making on climate change and adaptive land management.

Figure 5.2.1: SASSCAL Data and Information Portal

The SASSCAL Data and Information Portal is an open online environmental data and information portal that can be accessed freely using any web browser at http://data.sasscal.org (Figure 1). As a central data and information hub, the SASSCAL Data and Information Portal allows for the management, analysis, visualisation, linkage, and presentation of various types of resources, including time series data, geospatial data, documents, and others (Figure 2). Its advanced search functionality is supported by comprehensive metadata records for all resources that the system makes available. The system is fully interoperable and receives high-level acceptance among users from a wide user community, demonstrated by an average of 50,000 page impressions per month.

At the end of 2017, the SASSCAL Data and Information Portal contained data from 640 environmental measurement stations, including more than 700 hydro-climatic time series data records and more than 250 geospatial data sets from more than 70 regional and international organisations, as well as numerous documents. Data are added continuously. Resources can be

searched using keywords or temporal or spatial extent, and by means of predefined areas of interest, such as district boundaries or study sites.

5.2.1 Stakeholders

Implemented and operated by the SASSCAL Open Access Data Centre (OADC), the SASSCAL Data and Information Portal ensures that the research deliverables resulting from the SASSCAL 1.0 Research Portfolio are hosted and made available according to stakeholder demands. The portal offers a fine-grained user permission control approach which allows the data owner to upload and update data but also permits setting up access permissions.

Notably, the resources hosted by the SASSCAL Data and Information Portal are not limited to the SASSCAL research outputs, but also extend to publicly accessible data from other sources relevant to researchers and stakeholders, including the research community, decision makers and the public.

5.2.2 Systems Architecture and Functionalities

The SASSCAL Data and Information Portal is based exclusively on open source solutions, while ensuring data interoperability and allowing extensibility. The system is based on a three-tier architecture with user frontends and server functionality for database operations (Figure 3). All data are processed on the server, putting less strain on hardware capacity at the end user's side.

Following a fully open-source approach, the system builds on PostgreSQL/PostGIS databases for data management, an Apache HTTP Server for web services, and a Catalog Service for the Web (CSW) server for metadata representation, and implements the Bootstrap web framework with different JavaScript libraries to create a user-friendly and intuitive graphical user interface. The metadata model is based on ISO standards (e.g., ISO, 2005) and further adheres to specifications of gazetted metadata standards in the SASSCAL countries. A full description of the technical details of the SASSCAL Data and Information Portal can be found in Zander and Kralisch (2016).

In its current version, the SASSCAL Data and Information Portal offers a wide range of functionalities. Advanced gap analysis for time series data, visualisation, and manual and automated import/export tools for various data types have been implemented, as have sophisticated web mapping functions for geospatial data exploration. Geospatial data and metadata are provided through standardised web services.

5.2.3 Outlook

The SASSCAL Data and Information Portal architecture serves the SASSCAL objective in developing and operating a regional resource and data hub for southern Africa. Its current functionalities already ensure that it can host data and information from any relevant research project. To allow for the consideration of new user demands, the data portal will be continuously enhanced in the future. For example, it will cater to the integration of additional data processing and analysis tools; advanced hydrological, climate, and other environmental models; and offer a link to other SASSCAL data products, such as SASSCAL WeatherNet (www.sasscalweathernet.org) and the SASSCAL observations net (www.sasscalobservationnet.org). The integration of advanced filter and search tools, documentation, and online help functions will ensure a seamless and intuitive user experience.

The SASSCAL Data and Information Portal aims at providing open online data and information resources, but at the same time intends to protect the intellectual property rights of the scientific and research community. Providing user functions for data access, but also for uploading new data, it serves as a flexible one-stop solution for data management, data exchange, and dissemination of research results.

5.3 The ICOS Carbon Portal

The ICOS Carbon Portal (CP), https://www.icos-cp.eu, has been developed on the basis of the Carbon Portal white paper written in 2012. The white paper document is available on ICOS Alfresco. CP was developed as the one stop shop for all ICOS data products, i.e.: “a virtual data centre where ICOS data can be discovered, accessed and visualized, and where users can also deposit data products based on ICOS data”. The data system is integrated with a metadata system that describes the data and its provenance. Furthermore, CP provides capabilities for advanced web based services that provide researchers, general public and decision makers with useful higher level products based on ICOS data.

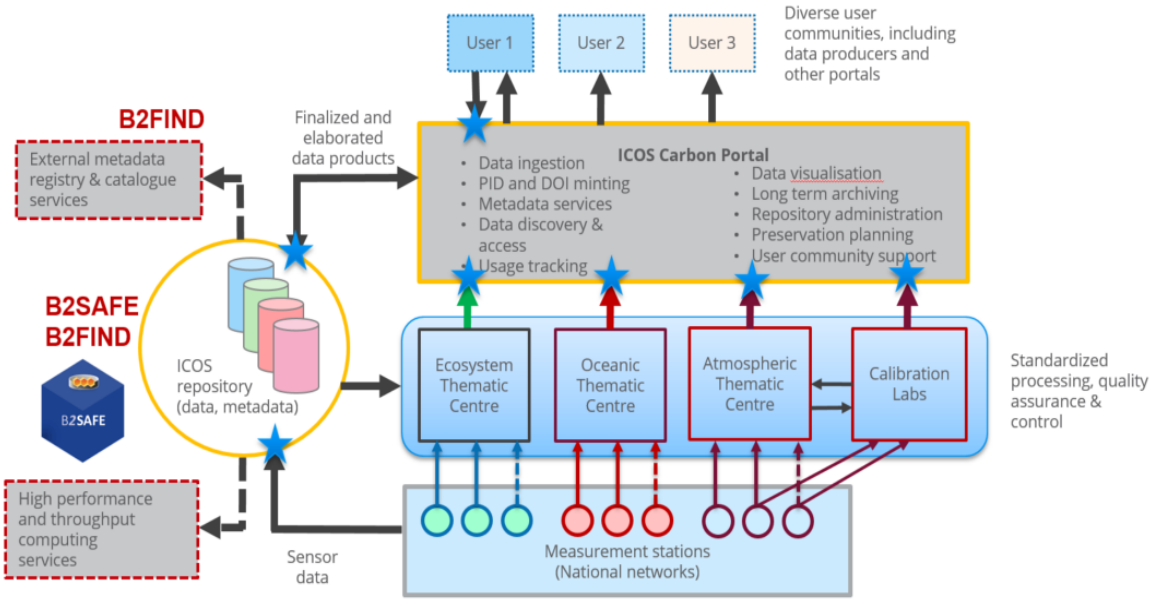


Figure 5.3.1 Simplified data flow within ICOS Research Infrastructure

The basic principles of the CP are data security, long term archiving through a trusted repository, enforcing the data policy and user friendly operation. As a service to the data providers, CP will keep track of the use of the data and its citation. By default, the CP supports machine to machine access to data and metadata. For human users CP adds user friendly web services on top for data discovery and access.

All ICOS data is open data, licensed under a Creative Commons International 4.0 Attribution (CC4BY).

All developments by CP are open source and are based on open source libraries and tools. The sources are licensed under GPL and are available from https://github.com/ICOS-Carbon-Portal. The backend skeleton Portal metadata and data services are generic and fully customizable and can be adapted to any project or look and feel. Landing pages can be stylized to an identity that is coupled to the data object type, so can depend on theme or data provider. All services in backend and frontend are dockerized and are fully scalable.

5.3.1 Carbon Portal Data Ingest

The philosophy of CP is to treat all data objects equal and preserve the complete integrity of all data objects, so the actual data is never touched or changed up to the bit level. This goes for all data levels, i.e. from raw data, NRT data, final data quality-controlled data up to elaborated data products. CP strives for the maximum granularity of Data Objects.

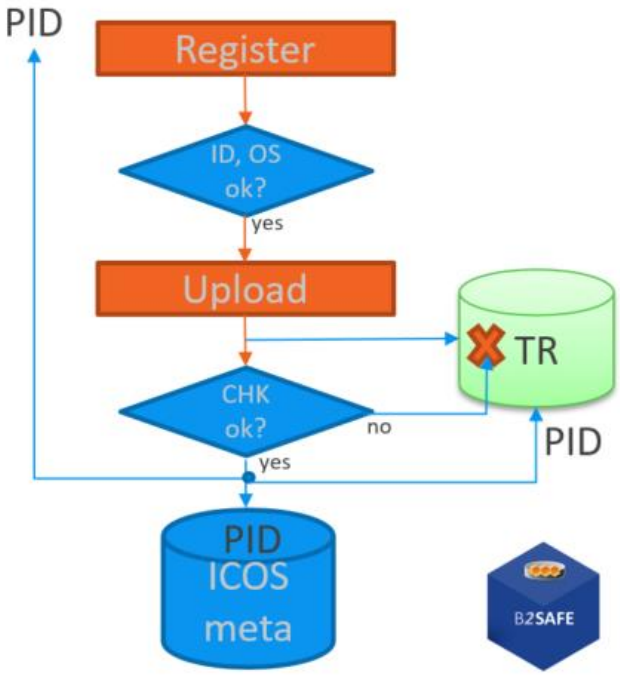


Figure 5.3.1 A simplified schematic of the ICOS central data ingest that enables robust, persistent identification and transparent and secure data ingest directly into the trusted repository (see also https://github.com/ICOS-Carbon-Portal/meta)

Before ingestion CP requires the uploader to calculate the SHA256 checksum of the data object. All ingestion data transport uses standard http(s) put and get methods, and can be invoked by for example using the curl program. In the first stage of ingestion the uploader informs through a small metadata packet in JSON format of the object specification and the checksum of the data object together with some minimal provenance metadata that informs on the uploader, the spatial and/or temporal coverage that the data relates to for as far as applicable and depending of the object specification also on other important information like station, measurement level and instrument ID. Only objects with a known and registered Object Specification type are accepted. After successfully registering in this first step the user can start uploading the data object. While the uploader streams the data to CP, the data is forked and streamed at the same time to the B2SAFE trusted repository.

When the object specification defines the data format of the file, a check is performed after the complete upload, to check the compliance to the data format and even possibly the validity of the data columns and spatial and temporal coverage as contained in the data file. Any deviation from the definition or prescribed metadata results in refusal of the file and abortion of the ingestion. The successful parsing of the data for text files also results also in the generation of binary CP-internal representations of the data that are used for the visualisation of time series in the data preview.

After upload completion, the checksum of the upload is compared with the registered checksum and when ok, a handle PID is minted for the data object and returned to the user. The metadata from the metadata packet is then added to the metadata repository and enriched with information on the PID, the checksum and other Object Specification dependent metadata. The suffix of the data object PID consists of the first 18 characters of the checksum of the data object and is thus unique

for the data object. Later the PID suffix can at any time be compared with the SHA256 checksum of the data object to ensure that the data is up to the bit and exact copy of the original data object.

5.3.2. The CP metadata system

The metadata that accompanies the data objects is maintained in a versioned so called RDF triple store, following the Web 3.0, linked open data approach. The database can be queried using an open SparQL endpoint at https://meta.icos-cp.eu/sparql. The metadata store fully supports date versioning and data collections. It is machine actionable through standard http(s) protocol. The metadata store is fully described by the underlying ontology, that again itself is defined in RDF through the OWL language.

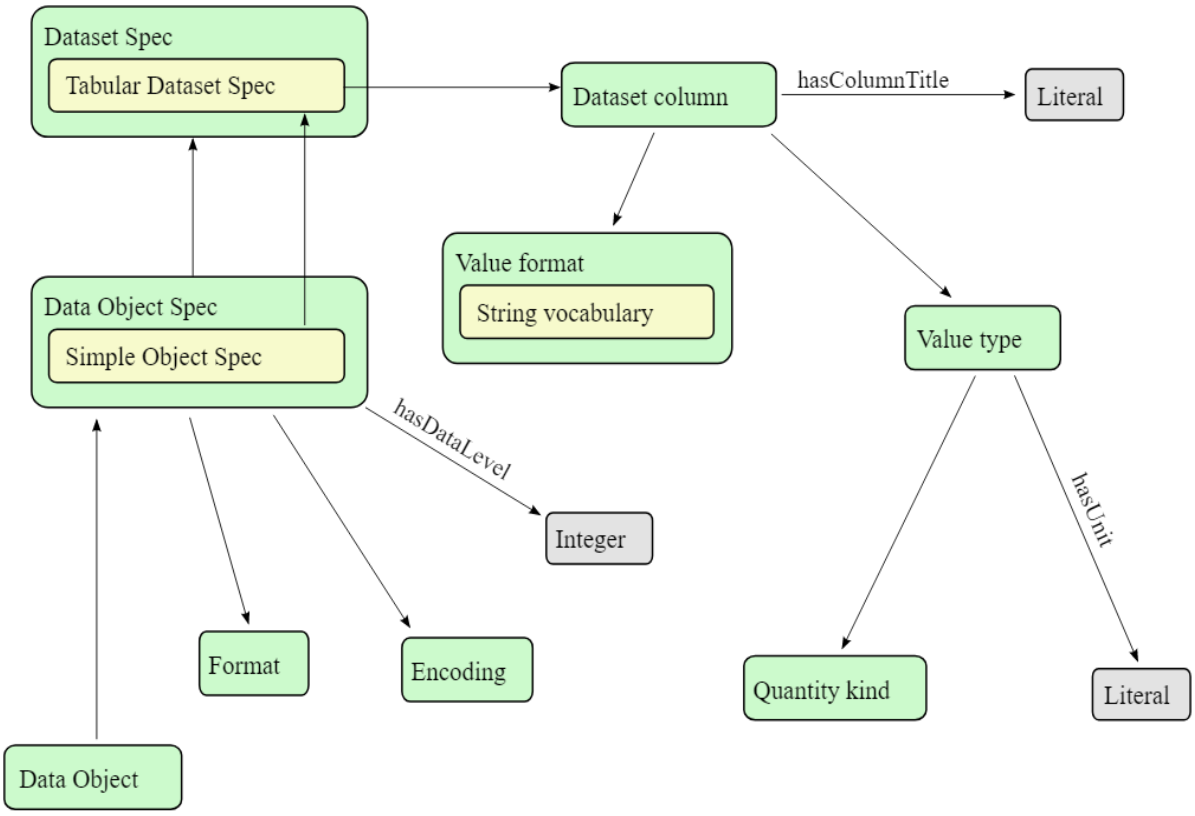


Figure 5.3.2 The simplest ICOS data object model for time series (see also https://github.com/ICOS-Carbon-Portal/meta)

The design of the metadata system is fully configurable to act with a single or multiple portal front ends using a single or multiple metadata stores. This means that for example multiple infrastructures can have their own differently styled data portal and use one single metadata store, or that one infrastructure has one portal that uses several external metadata stores, or that several infrastructures use one common portal that relies on a set of federated metadata stores, one per infrastructure. All completely transparent to the outside user.

The ICOS CP metadata store is for example shared with the Swedish SITES national infrastructure that has its own dedicated and styled portal, while ICOS Sweden is just using the metadata store backend and data is served through the ICOS CP portal.

The metadata system supports versioning of data, dynamically growing data and collections.

5.3.3 CP data discovery

The main entry point for data discovery for humans is https://data.icos-cp.eu. Here a set of filters can be easily set to filter to the data sets that the user actually is looking for. The list of data objects that fulfils the set of filters is display dynamically. Changing the filters also dynamically updates the remaining options for the other filters that comply with the other filter settings. Filters can and will be added, removed and applied incrementally. From the results page the user can view the most relevant information on the data object and/or drill down to the data object landing page for all relevant metadata. Most data objects can be previewed, see data visualisation. Most data objects can also be added to the user's data cart for easy download, see data access.

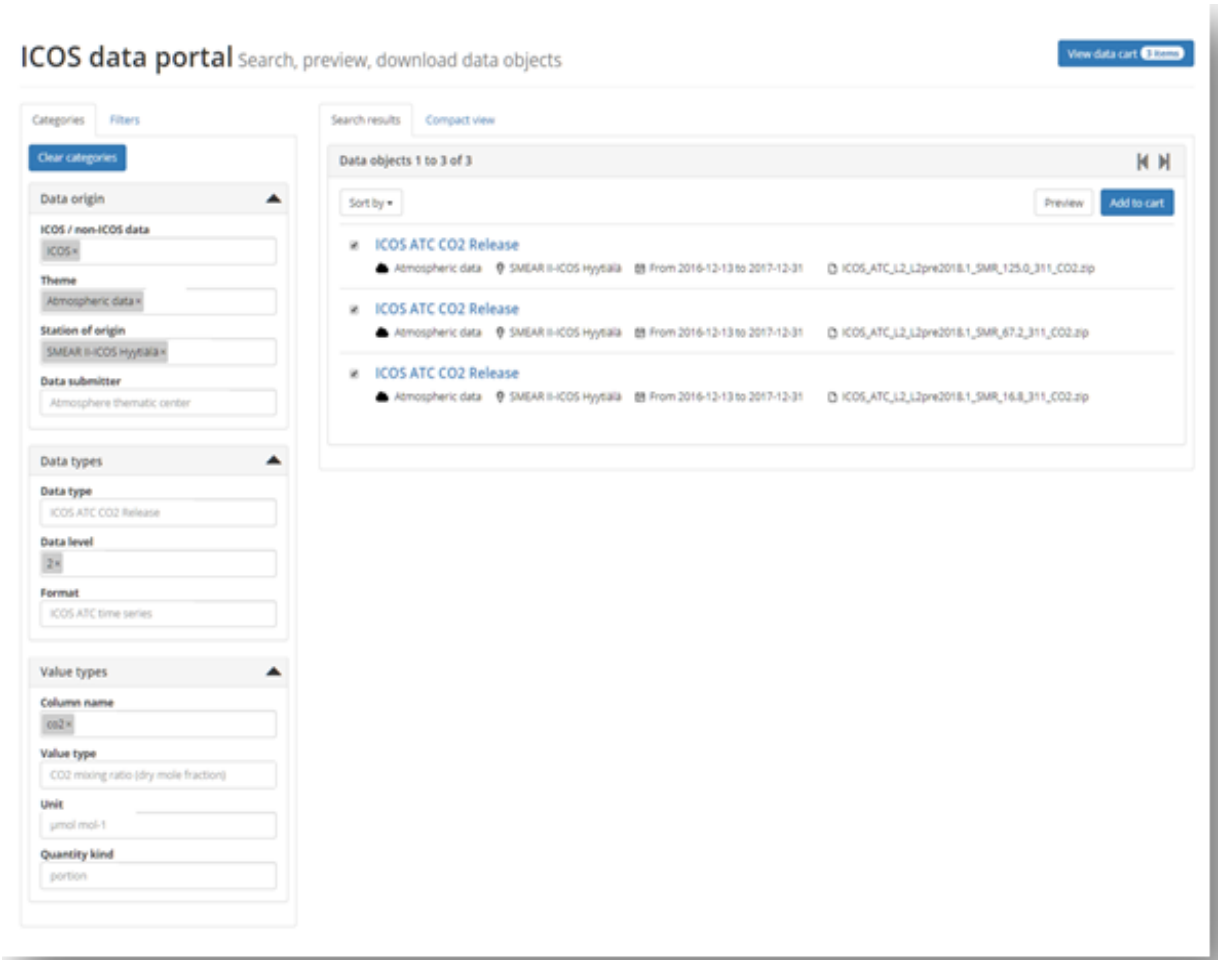


Figure 5.3.3 Example of the ICOS Data Portal Search Results

5.3.4. CP data access

Data access is provided through the PID (or DOI) of the data objects. Resolving this PID through the handle or DOI system leads normally to a landing page that contains a link to the data object(s). In case of non-ICOS data objects this link can point to another data portal due to data license restrictions. Raw data objects are currently also not directly downloadable but require contact with the relevant thematic centre.

The data discovery tool allows to add selected data objects to the user's data cart from where the collected objects can be downloaded in one batch into a single zip archive.

All data downloads are logged and ICOS data has a data licence check implemented before the download to inform the user of the ICOS CC4BY licence and its implications. Users can easily track the number of downloads per dataset, country, station, contributor and/or theme, categorized by time and country of the download.

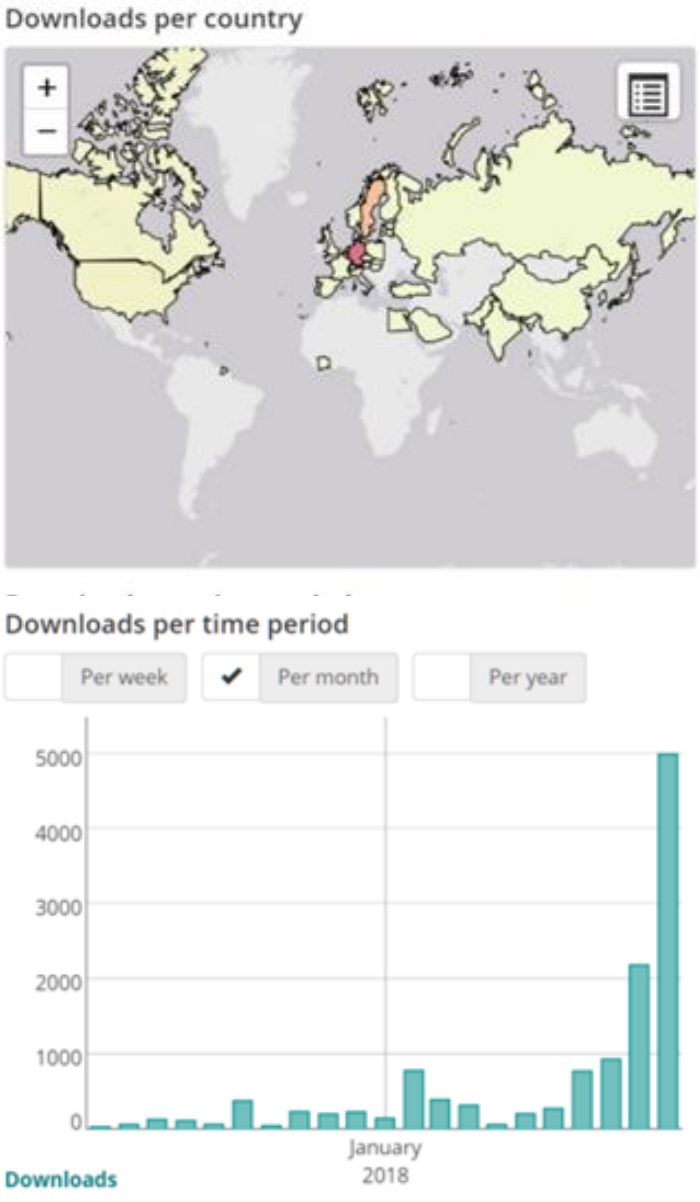


Figure 5.3.4 Example of Download Metrics

5.3.5 CP data visualisation

Time series and spatial data sets can be previewed directly from the data portal in the search results and in the data cart for a quick check just before download. The visualization supports the overlaying to append time series for a single column and the overlaying of overlapping time series of from different stations, instruments and/or measurement heights. A fully interactive map or chart is shown that can be reproduced in any web portal or page in an iframe by using the provided link.

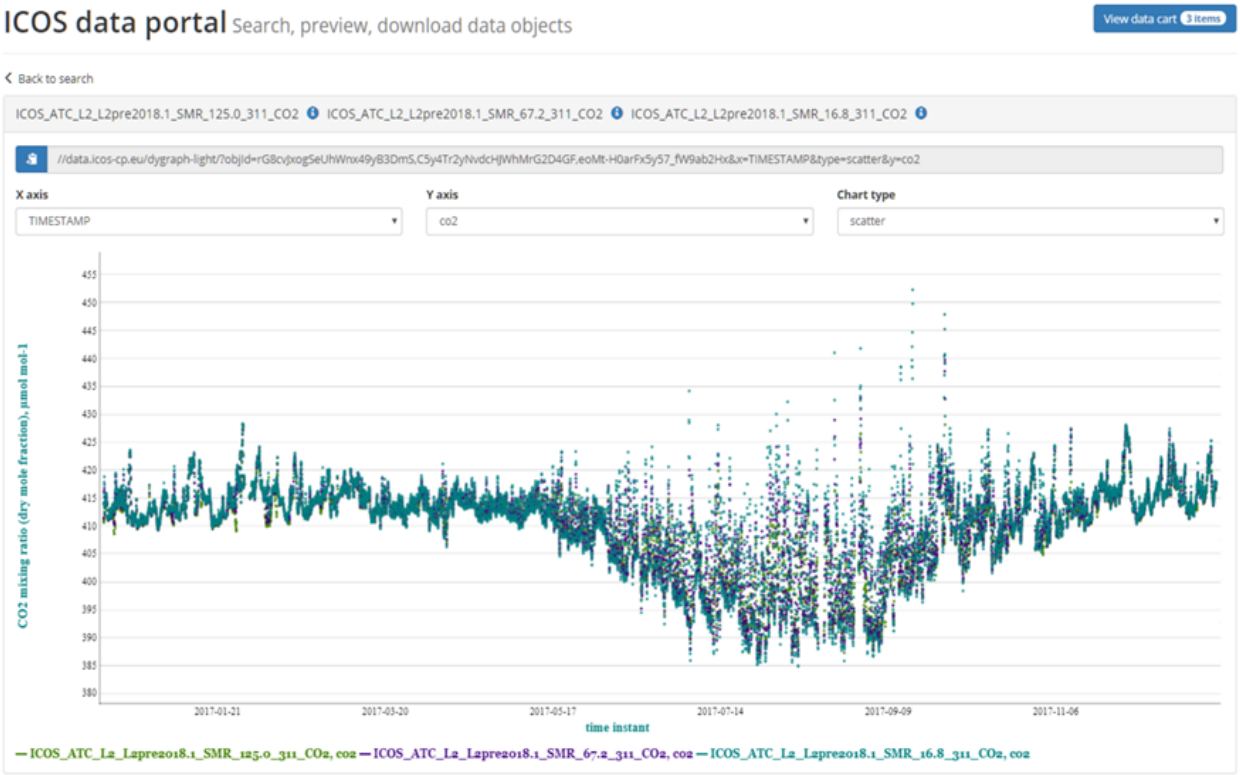


Figure 5.3.5 Example of Data Visualisation

An interactive tool allows to link atmospheric footprint data with modelled and measured time series at https://stilt.icos-cp.eu/viewer.

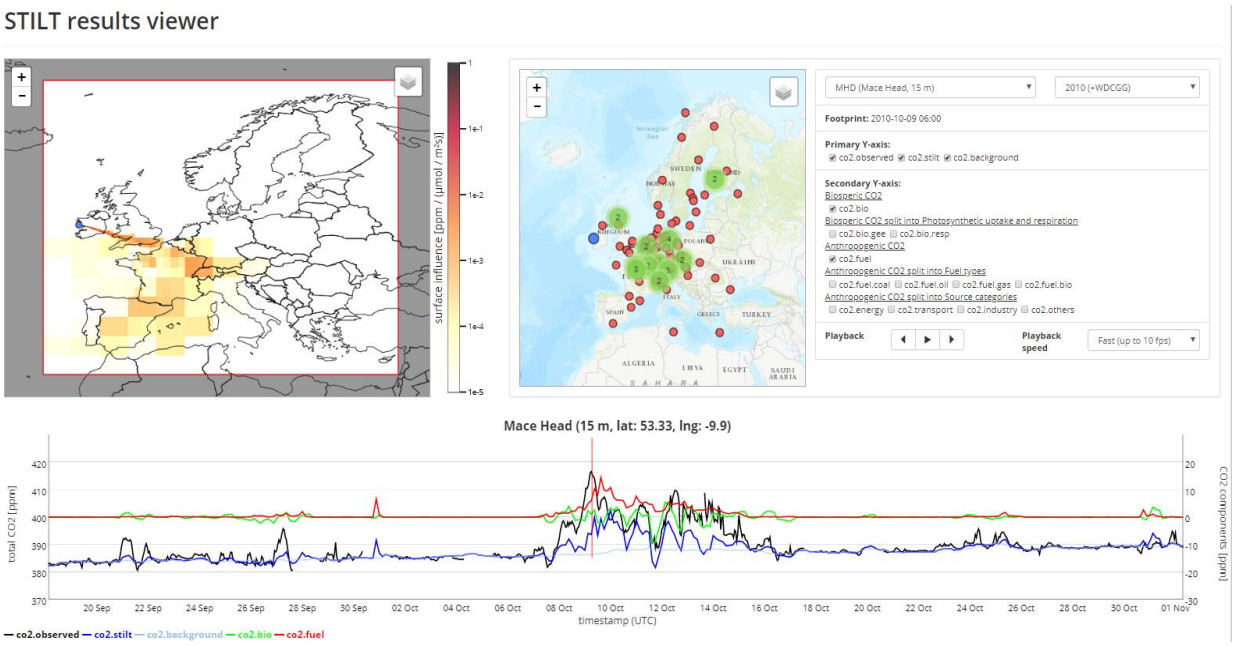


Figure 5.3.6 STILT Results Viewer

5.3.6. Virtual Research Environments

Carbon Portal offers scientists access to Jupyter notebooks that give direct access to the data objects stored at the ICOS CP. These notebooks are either run directly on the CP servers or are operated as virtual machines in the cloud, making use of a changeable amount of memory and cores for even the most demanding analyses. Notebooks can be shared among colleagues for collaborative analysis of for example model ensembles, sharing common input and output data and modelling resources.

For less advanced users that would have difficulties with programming, CP plans to provide interactive tools that give access to powerful models and data analysis tools. One example is the Stilt footprint calculator that allows users to perform footprint calculations using the Stilt Lagrangian footprint model for any point in Europe and period within the provided range in space and time. The results are immediately after calculation available in the Stilt results viewer and for download, together with the forward prognosis of CO2 concentrations at the chosen receptor point.

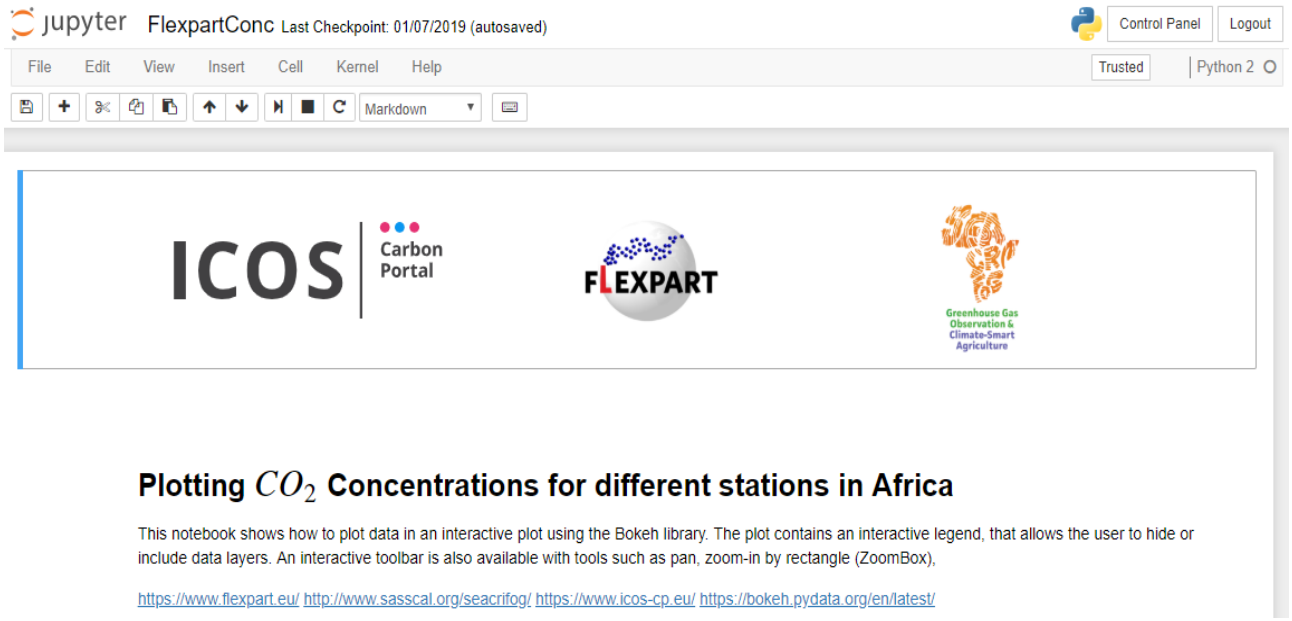


Figure 5.3.7 Incorporating ICOS Data into Jupyter Notebook

6 SEACRIFOG-Specific Requirements

6.1 Scope of Variables

A set of 58 environmental variables was identified by SEACRIFOG (see Deliverable report 4.1) to be essential for the systematic observation and characterization of the climate system in the context of the African continent and the surrounding oceans. SEACRIFOG Deliverables 3.1 and 4.2 specified the observational requirements for these essential variables, which thus also need to be met by respective data products.

The main requirements considered include spatial and temporal resolution as well as required accuracy (or maximum uncertainty, respectively). Various essential variables are part of the Essential Climate Variables (ECV), for which global requirements are already defined by the Global Climate Observing System (GCOS). The ECVs include some of the Essential Ocean Variables (EOV), for which observational parameters and requirements are defined under the Framework for Ocean Observing of the Global Ocean Observing System (GOOS). Compatibility goals for the in-situ observation of atmospheric variables are further defined by the WMO Global Atmosphere Watch (GAW). These global requirements were further refined and complemented by the SEACRIFOG consortium based on expert judgement in the context of the African continent.

The concept of data families and the community standards associated with each is discussed in detail in Mirtl et al (2018), and is reviewed in Annexure B.1.

These variables are listed and assessed in terms of their data family and data velocity/ volume implications in Annexure B.2. In summary, the following can be determined from the assessment of these variables:

6.1.1 Big Data

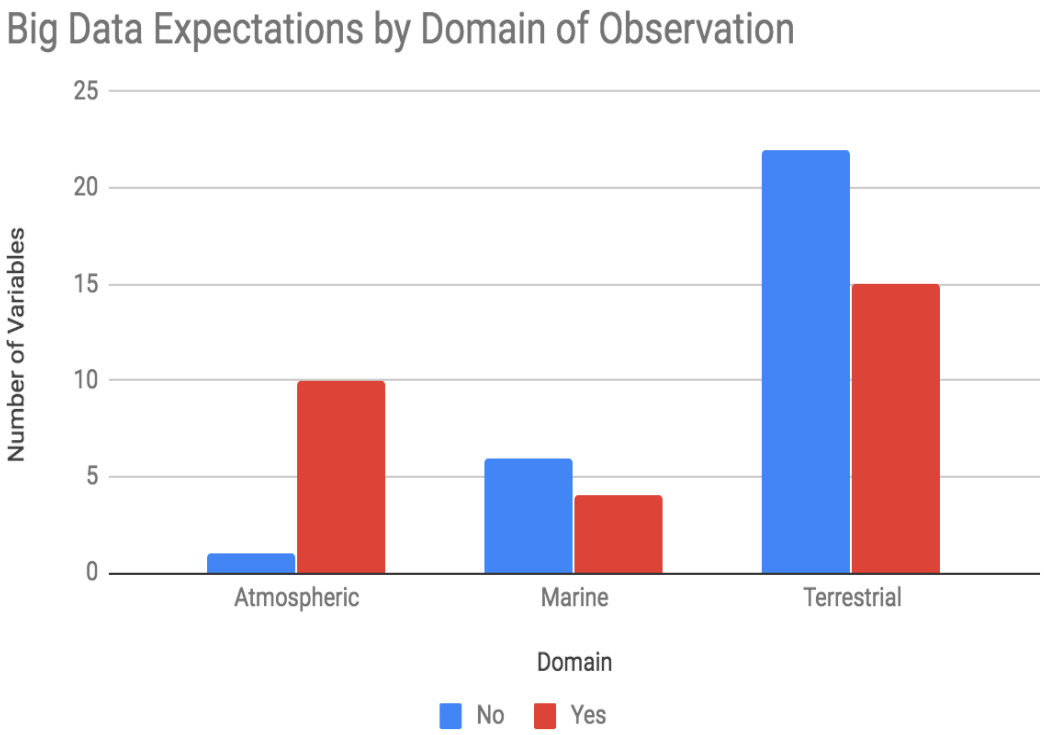


Figure 6.1.1. Expected Big Data Requirements per Domain. Yes: Big Data volumes and velocity expected due to fine spatial and/ or temporal resolution.

Firstly, we evaluated the big data implications of each variable in respect of temporal and spatial resolution. Variables with high frequency sampling rates (hourly or less) have high velocity, and require special measures in respect of processing, archiving, and data volumes. Frequent sampling and/ or fine-scale spatial resolution (100m or less) will, in combination, result in large datasets.

Design directives were derived from this analysis:

1. Most atmospheric observation data, and a large proportion of the terrestrial observation data will likely be 'Big Data', and as such, will be best provisioned in cloud services managed by a centralised infrastructure node. Given that many of the variables are remotely sensed (see below) this will make a lot of sense, and address some of the infrastructure and federation concerns expressed in other sections.
2. Many marine data sets will not be linked to a specific country if it is outside the EEZ, and could be hosted centrally too.

6.1.2 Observation Techniques

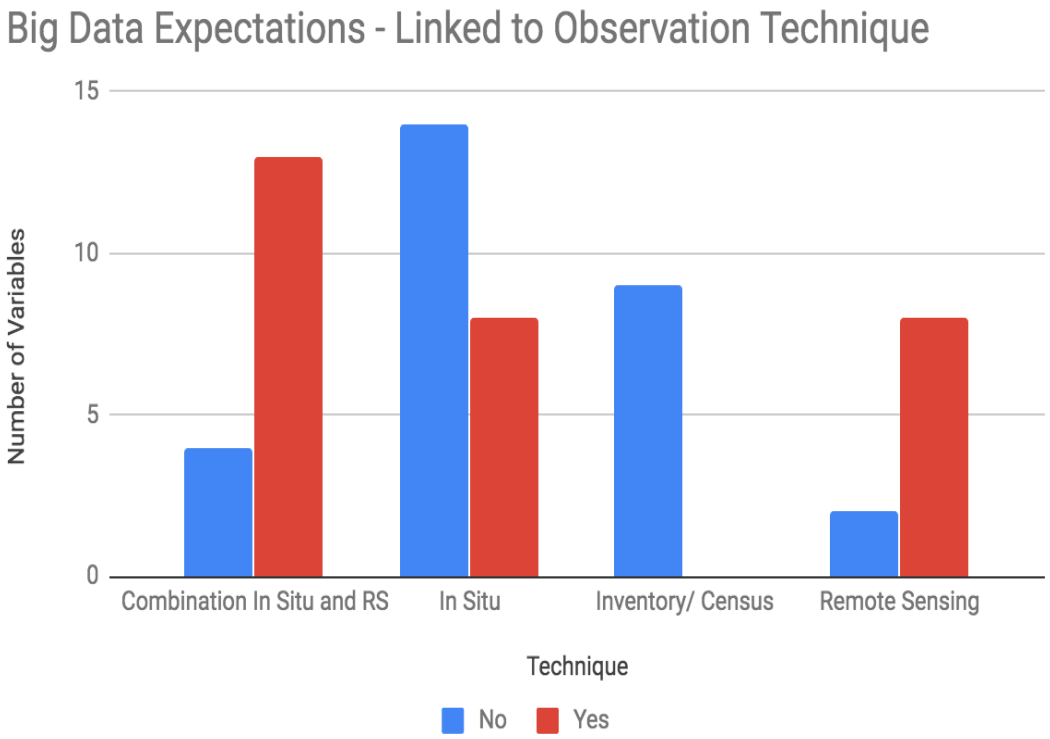


Figure 6.1.2 Expected Big Data Requirements by Observation Technique. Yes: Big Data volumes and velocity expected due to fine spatial and/ or temporal resolution.

The majority of big data volumes/ velocity expected derives from remotely sensed data at high spatial (sub 100m) and temporal (hourly or less) resolution, or from spatially disperse in situ observation that occurs at very high frequency (for example flux tower measurements).

From this assessment, it appears that at least the remotely sensed data, and possibly high velocity in situ data, should be considered for centralised management and preservation. This will avoid having to create relatively expensive 'big data' repositories in a large number of localities.

6.1.3 Data Families

In broad terms, there are five data families of interest to the SEACRIFOG consortium in respect of data infrastructure. These are reviewed in detail in Annexure B.1. We summarise them here:

1. **"Spatial" Data**: These are traditional spatial data sets (vector and raster data sets) that are relatively sparse in time, and continuous or near-continuous in space.
2. **"Time Series" Data:** Data that is (near) continuous in time and relatively sparse or discrete in space.
3. **"Cube" Data":** Multidimensional data, (near) continuous in time and space.

4. **“Structured” Data**: Tabular data or relational databases that do not fit one of the other categories.
5. **“Object” Data**: Digital objects that are not structured or contains indirect structures, such as reports, articles, media files, and similar.

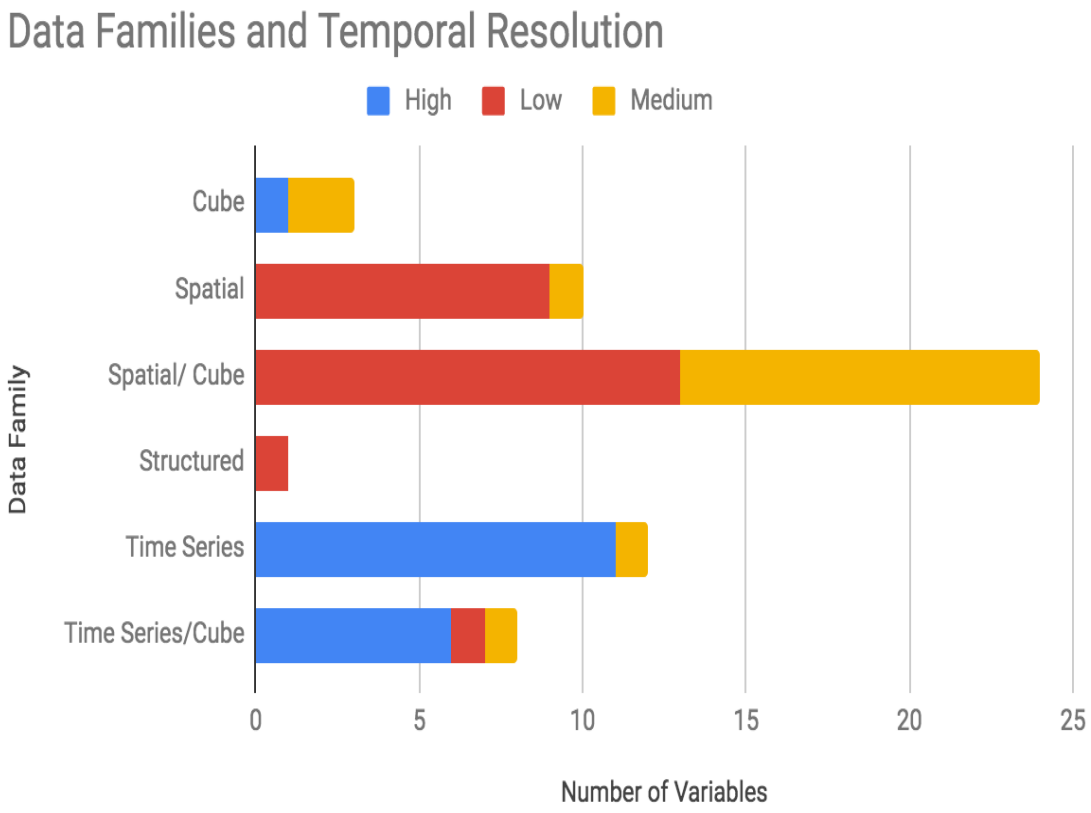


Figure 6.1.3. Data Families and Temporal Resolution.

Some datasets can be stored in more than one standardised datastore. Typically, this involves spatial data (stored as traditional spatial data or as a datacube), or time series data (stored as time series in relational tables, or as a datacube). We propose that the data velocity (frequency of update) be used a deciding factor in respect of which data family to select.

Our design considerations derived from the above can be summarised in the following table:

	Temporal Resolution (Velocity)		
	High | < 1 day	Medium | 1 day - 1 year	Low | > 1 year

Spatial Resolution	High | < 100m,	Cube Data	Cube Data	Time Series/ Spatial/ Cube Data
	Medium | 100m - 1 km	Cube Data	Time Series/ Spatial/ Cube Data	Time Series/ Spatial
	Low | > 1 km	Time Series/ Spatial/ Cube Data	Time Series/ Spatial	Time Series/ Spatial

One should also consider the practicality of data pipeline management implicit in the permutations above. Amongst other things, one should aim to

1. Store and process large remotely sensed datasets as close as possible to source, preferably in cloud-based services with processing and service publication capabilities.
2. For in situ observations with high velocity, provide centralised facilities in cases where it is not practical to host data infrastructure close to source, and invest in connectivity to enable transfer of high velocity data to a shared facility for processing and hosting.
3. For any other observations, provide centralised 'infrastructure of last resort' - but allow federated data hosting with centralised metadata aggregation and management.

Our assessment is summarised in the chart below.

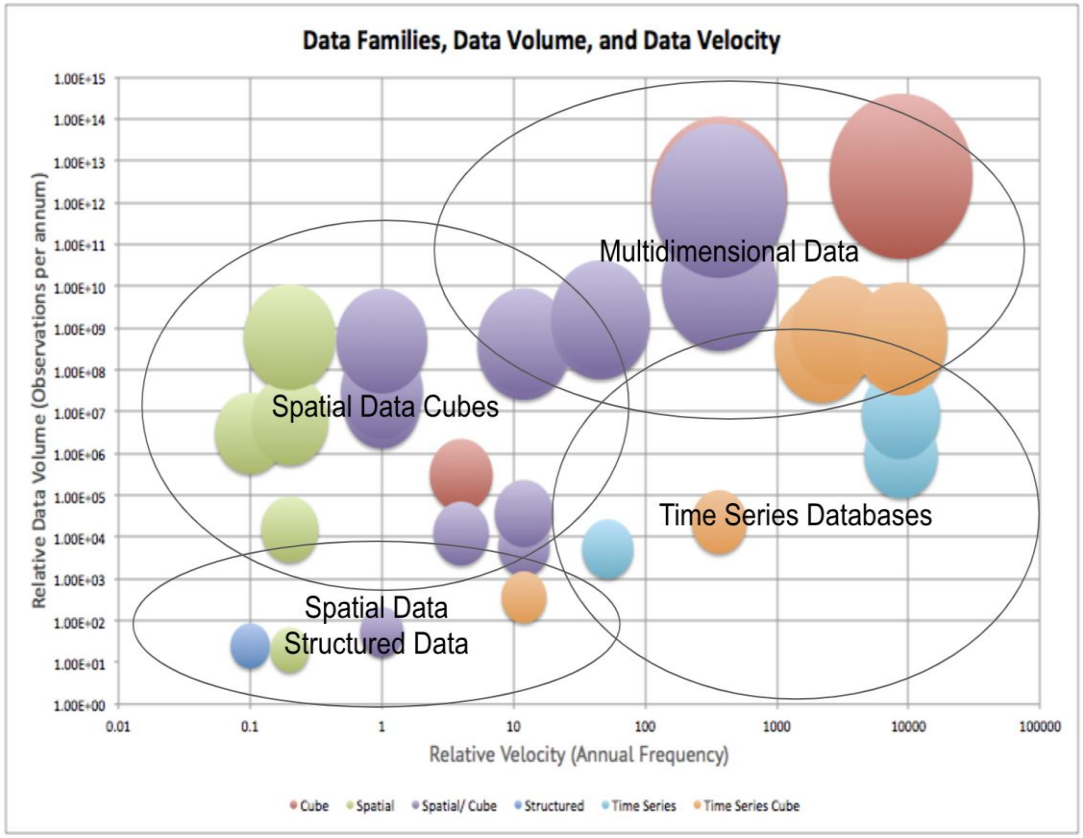


Figure 6.1.4. Data Family Assessment.

6.2 Current and Planned Observation Infrastructure

The SEACRIFOG project will determine an optimal distribution of observation infrastructure for GHG emissions by inverting a spatially-explicit GHG emission, transport and chemistry model. The purpose of this optimised distribution is to minimise uncertainty for the lowest possible investment costs.

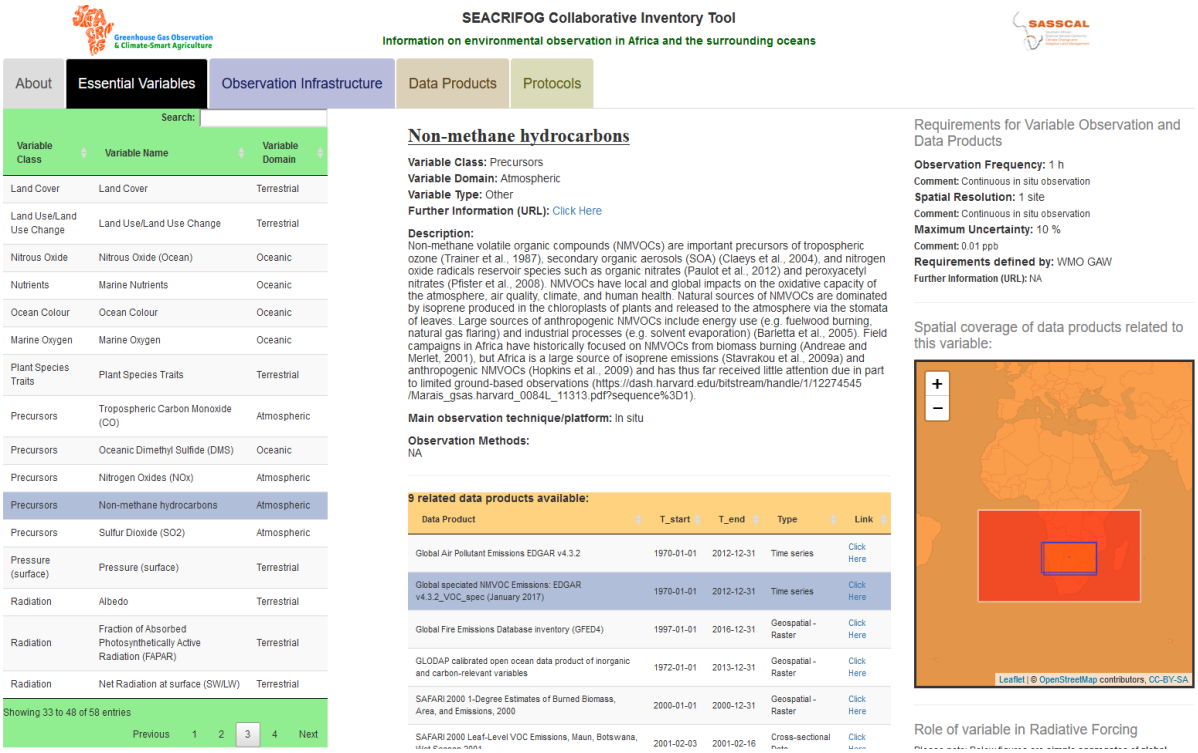
In addition, the project also identified suitable protocols for the measurement of the variables that were identified. These have to be linked to and incorporated into any data infrastructure that is implemented, since the protocol used for observation determines the extent to which data can be collated automatically, with or without adjustments. These considerations are discussed in more detail below (Section 6.3) and in Annexure A.5

6.3 The SEACRIFOG Collaborative Inventory Tool

The SEACRIFOG Collaborative Inventory Tool (https://seacrifog-tool.sasscal.org/) is a web-based application specifically developed by SASSCAL in line with SEACRIFOG, which serves to systematically capture, contextualize and visualize information and metadata on

- the essential variables identified by SEACRIFOG,
- existing and planned observation infrastructures in and around the African continent relevant to these variables,
- existing data products related to the essential variables
- existing methodological protocols relevant to the observation, data processing and modelling of the essential variables.

The tool further serves as a public resource, informing about the state of environmental observation across the African continent and the surrounding oceans and supporting research infrastructure development in line with the SEACRIFOG project.



The major functionality is the provision of comprehensive contextualized information and data for each essential variable with regards to corresponding observation on the African continent. When selecting a certain essential variable, the user is given access to all variable-specific information as well as all related observation infrastructures, data products and protocols stored in the database.

The tool was developed exclusively based on open source components. It was written in R using the Shiny package and is hosted on a cloud server rented by SASSCAL using 'Shiny Server Community version'. The data is stored in a relational PostgreSQL database. Read access is public whereas write access is limited to registered users.

The tool grew into its present form rather organically and through various iterations, mainly with the purpose to structure the work of SASSCAL and the project partners involved in SEACRIFOG work packages 3 and 4 and to facilitate the technical and scientific input of a wide range of collaborators. Given this specific purpose, it constitutes a rather basic tool (both in terms of scope and performance) compared to the more advanced portals presented in this section. However, it has proved to be very helpful and valuable to various SEACRIFOG tasks. Furthermore, since it was developed specifically for the SEACRIFOG project, the tool can be considered as a first iteration towards the development of a more advanced prototype of an African e-infrastructure for environmental data in line with SEACRIFOG WP5 and hold some valuable insights and lessons.

Additional functionality and content, which may/should feature in the prototype e-infrastructure includes

- metadata (catalogue) import/export
- advanced search functionality, allowing to search by more attributes
- observation site-level instrumentation inventory and corresponding links to essential variables observed

- ability for adding, editing and deleting individual observation sites
- expansion of space-borne mission and instrument inventory and corresponding links to essential variables
- visualization of temporal coverage of data products for each essential variable
- automated comparison of data product metadata against essential variable requirements

The vision for the tool was to provide the digital infrastructure for a community of collaborators (the SEACRIFOG consortium as well as the wider environmental research community) with the aim to continuously pool their collective input to compile inventories of relevant observation infrastructures, data products and protocols that are as complete and up-to-date as possible. While there has been significant input from collaborators, the present inventories are far from exhaustive. In order to scale up corresponding efforts and outcomes, besides maximising the web-performance and user friendliness, more emphasis would need to be placed on the establishment of a community of practice (i.e. active users from the environmental research community) who at the same time contribute to and benefit from the system.

7 Conclusion

Previous work carried out by the SEACRIFOG consortium creates a good basis for data portal design. Descriptions regarding data availability, widely accepted measurement protocols and metadata standards in WP4 and WP5 are useful when considering requirements and solutions for technical, practical and licensing aspects. Several examples of existing data portal infrastructures in this report frame clearly variety of available technical solutions and different ways to build services for scientific community. SEACRIFOG project and Collaborative Inventory Tool has managed to define the most important Essential Climate Variables needed for evidence-based decision making and environmental monitoring in African context. However, service catalogue describing data availability, although based on co-operation of different data providers, cannot be directly constitute a basis for African Research Infrastructure.

Firstly, data products listed in service catalogue are not products deriving from the work that SEACRIFOG consortium has carried out. They are in most cases data products or research outputs deriving from remote sensing community or project funded research programs. Therefore, data products listed in collaborative tool cannot be considered as data deriving from formal RI. Secondly, if formal Research Infrastructure will be built, a formal and legally valid administration will be needed, including governance and guidance structures (e.g. GA: general assembly and SAB: scientific advisory board). Administration of African RI should define and approve data policy practises, measurement methods and platforms, metadata formats and technical facilities that are dedicated for operational data portal. Thirdly, scientific community will benefit from services provided by the African RI, in case platforms and application services are designed to support value added data produces that are deriving from the work carried out by the scientific community. This is also a way to integrate stakeholder communities, Research Infrastructure and scientific community providing information for evidence-based decision making.

Previous experiences from operational data portals and European RI's creates a good basis for technical implementation plan and blueprint data infrastructure in SEACRIFOG project. However,

African RI may not be able to harmonize and standardize GHG measurements in such extent that for example ICOS. Local existing research environments and existing measurement techniques may require solutions for several instrumentation setups. However, some parts of the raw sensor data processing, post processing and data repository solutions can be automated using similar principles than used in ICOS Carbon Portal. Technical solutions implemented in ICOS CP are available through GitHub and can be modified and adjusted for blueprint data infrastructure in Deliverable 5.4.

Literature

Beck, Johannes & López-Ballesteros, Ana & Omar, Abdirahman & Johannessen, Truls & Skjelvan, Ingunn & Helmschrot, Jörg & Saunders, Matthew. (2018). SEACRIFOG Deliverable 4.1: Identification of Key Variables for Climate Change Observation Across Africa. 10.13140/RG.2.2.28727.98724. https://www.seacrifog.eu/fileadmin/seacrifog/Deliverables/2018.08.18_SEACRIFOG_Deliverable_4.1_doi.pdf

Beck, Johannes & López-Ballesteros, Ana & Skjelvan, Ingunn & Bob Scholes, Robert & Vermeulen, Alex & Helmschrot, Jörg & Saunders, Matthew. (2019). Africa-Wide Climate System Observations: Data Requirements and Availability (SEACRIFOG Deliverable 4.2). 10.13140/RG.2.2.18543.28320. https://www.seacrifog.eu/fileadmin/seacrifog/Deliverables/2019.01.31_SEACRIFOG_Deliverable_4.2_doi.pdf

Benecke, C., Hartje, H., von Sachsen-Coburg und Gotha, V. und Spils ad Wilken, H. (2014): Der konsolidierte Jahresabschluss. Transparenz für Unternehmer und Banken. Frankfurt am Main. Internetseite DLG-Verlag GmbH

BMEL, Bundesministerium für Ernährung und Landwirtschaft (2014): Grundzüge der Gemeinsamen Agrarpolitik (GAP) und ihrer Umsetzung in Deutschland. http://www.bmel.de/DE/Landwirtschaft/Agrarpolitik/_Texte/GAP-NationaleUmsetzung.html. Stand 10.2.2015

BRH, Bundesrechnungshof (2012): Bericht nach § 99 BHO zur Gewinnermittlung nach Durchschnittssätzen bei land- und forstwirtschaftlichen Einkünften (§ 13a Einkommenssteuergesetz)

CTS (2016): Core Trustworthy Data Repositories Requirements, https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf [Last accessed 2019-02-02]

Downs, Robert [2017]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-3/W2, 2017. 37th International Symposium on Remote Sensing of Environment, 8–12 May 2017, Tshwane, South Africa https://doi.org/10.5194/isprs-archives-XLII-3-W2-51-2017

Fiore, N., Rosati, I., & Tagliolato, P., and Oggioni, A. (2015). The LifeWatch e-infrastructure for biodiversity and Ecosystem research.

GEO (2015): Data Management Principles (DMP) Implementation Guidelines: Life-Cycle Data Management Principles. 2015. Group on Earth Observations. https://www.earthobservations.org/documents/geo_xii/GEO-XII_10_Data%20Management%20Principles%20Implementation%20Guidelines.pdf [Last accessed 2019-02-02]

Alex R. Hardisty, William K. Michener, Donat Agosti, Enrique Alonso García, Lucy Bastin, Lee Belbin, Anne Bowser, Pier Luigi Buttigieg, Dora A.L. Canhos, Willi Egloff, Renato De Giovanni, Rui Figueira, Quentin Groom, Robert P. Guralnick, Donald Hobern, Wim Hugo, Dimitris Koureas, Ji Liqiang, Wouter Los, Jeffrey Manuel, David Manset, Jorrit Poelen, Hannu Saarenmaa, Dmitry Schigel, Paul F. Uhlir, W. Daniel Kissling: THE BARI MANIFESTO: AN INTEROPERABILITY FRAMEWORK FOR ESSENTIAL BIODIVERSITY VARIABLES,
Ecological Informatics, 2018, ISSN 1574-9541

https://doi.org/10.1016/j.ecoinf.2018.11.003
http://www.sciencedirect.com/science/article/pii/S1574954118301961

Hugo, W. Guidance on Data Policy and Supporting Open Licenses, SEACRIFOG Deliverable 5.2 (In Process)

Ana López-Ballesteros, Johannes Beck, Antonio Bombelli, Elisa Grieco, Eliška Krkoška Lorencová, Lutz Merbold, Christian Brümmer, Wim Hugo, Robert Scholes, David Vačkář, 2018: TOWARDS A FEASIBLE AND REPRESENTATIVE PAN-AFRICAN RESEARCH INFRASTRUCTURE NETWORK FOR GHG OBSERVATIONS, Environmental Research Letters, Volume 13, Number 8, http://iopscience.iop.org/article/10.1088/1748-9326/aad66c/meta

López-Ballesteros, Ana & Beck, Johannes & Saunders, Matthew. (2019). Methodological protocols for the observation of climate change across Africa: an assessment of the current approaches with insights into the feasibility of implementation. 10.13140/RG.2.2.29998.97607. https://www.seacrifog.eu/fileadmin/seacrifog/Deliverables/2019.01.31_SEACRIFOG_Deliverable_4.2_doi.pdf

Mirtl, M., E. Borer, I. Djukic, M. Forsius, H. Haubold, W. Hugo, J. Jourdan, D. Lindenmayer, W.H. McDowell, H. Muraoka, D. Orenstein, J. Pauw, J. Peterseil, H. Shibata, C. Wohner, X. Yu & P. Haase, 2018: Genesis, goals and achievements of Long-Term Ecological Research at the global scale: A critical review of ILTER and future directions. Science of the Total Environment. Ref.Nr. STOTEN-D-17-06809.

OECD (2017): H2020 Programme - Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020, Version 3.2, 21 March 2017. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf [Last accessed 2019-02-01]

Rauber, A., Asmi, A., van Uytvanck, D., Proell, S (2016): Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). doi: 10.15497/RDA00016

Uhlir, Paul F. (editor) (2016): LEGAL INTEROPERABILITY OF RESEARCH DATA: PRINCIPLES AND IMPLEMENTATION GUIDELINES RDA-CODATA Legal Interoperability Interest Group September 8, 2016, http://www.codata.org/uploads/Legal%20Interoperability%20Principles%20and%20Implementation%20Guidelines_Final2.pdf [Last accessed 2019-02-01]

WDS (2015): Data Sharing Principles, http://dx.doi.org/10.5281/zenodo.34354 [Last accessed 2019-02-04]

Zander, F. & Kralisch, S. (2016) River Basin Information System: Open environmental data management for research and decision making. ISPRS International Journal of Geo-Information, 5, 123. doi:10.3390/ijgi5070123.

Annex

AnnexAnnexure A. Detailed Design Considerations

The design considerations derived from this report are listed in this annexureannexureannex. Each consideration is classified as follows:

1. **Aspect**: the major grouping or consideration class that is addressed.
2. **Architecture**: we foresee the development and specification of several 'architectures' in the Blueprint for Data Infrastructure (Deliverable 5.4). In this list, we identify the following architecture groups:
    a. PG: Policy and Governance
    b. SS: Standards and Specifications
    c. SI: Software Choices and Implementations
    d. HN: Hardware and Network Choices
    e. GC: Soft Infrastructure (guidance, procedures, capacity building)
3. **Sources and References**: These include all of the sources discussed in the body of the document, from section 2 to 6. If a consideration is encountered in more than one section, it is included at its first appearance but cross-referenced to other sections.

A.1 General Considerations

These requirements are based on the concerns determined from Section 2 - largely focusing on ENVRI Plus and FAIR.

#	Aspect	Design Consideration	Architecture	Source or Reference
1.1	Findable	Metadata must be available for all digital objects, datasets, and dynamically updated datasets	PG	2: FAIR, ENVRI 4: WDS, GEO, ILTER
1.2	Findable	Metadata scope must include citation and discoverability elements (for example spatial, temporal, and topic coverages)	SS	2: FAIR, ENVRI
1.3	Findable	Persistent identifiers, such as DOIs, are preferred for unique reference and unambiguous citation	SS	2: FAIR 4: RDA, GEO, WDS, ILTER
1.4	Findable	SEACRIFOG to agree minimum metadata standards for each identified data family	PG, SS	2: FAIR 6: SEACRIFOG

1.5	Findable	Metadata is required for support elements (protocols, instrumentation, sites and platforms., initiatives, variables, …)	PG, SS	6: SEACRIFOG 4: ILTER
1.6	Accessible	Use open licenses where appropriate	PG	2: FAIR 4: WDS, GEO, ILTER
1.7	Accessible	Maintain access via persistent identifiers - digital objects and data must always be available via metadata links	SS	2: FAIR 4: RDA, GEO, WDS, ILTER
1.8	Accessible	Metadata and licenses need to be machine readable	PG, SS	2: FAIR 4: RDA, GEO, WDS, ILTER
1.9	Interoperable	Both metadata and data need to be available in generally agreed standard formats, services syntax, and semantic alignment.	SS	2: FAIR 4: GEO, ILTER, WDS
1.10	Re-Usable	Links to code and algorithms should be made available to allow reproducibility	SS	2: FAIR 4: RDA, WDS, ILTER
1.11	Re-Usable	Metadata focused on re-use should be made available, and be based on notions of data value chains.	SS	2: FAIR 4: RDA, GEO, GEO-BON, ILTER

A.2 Africa-Specific Considerations

These requirements are based on the concerns determined from Section 3.

#	Aspect	Design Consideration	Architecture	Source or Reference
2.1	Policy: Collaboration	Collaboration will be required across national boundaries in Africa, requiring policies and collaboration agreements within the SEACRIFOG consortium, with collaborators and service providers, and for participants.	PG	3. AFRICA 6. SEACRIFOG
2.2	Governance: Administration	A secretariat and administrative capacity will be required to coordinate a continent-wide RI	PG	3. AFRICA 6. SEACRIFOG
2.3	Governance: Funding	Funding (capital and operations) will be required in respect of 1. Secretariat and administration	PG	3. AFRICA 6. SEACRIFOG

		2. Observation network and infrastructure 3. Data infrastructure 4. Capacity building, collaboration, workshops, and network events Costs in Africa are likely to be higher than in developed countries, possibly with the exception of lower-skilled employee costs.		
2.4	Governance: Impediments	Inter-country travel in Africa is not simple: visas are difficult to arrange, some destinations are not safe, and direct flights are often not possible.	PG	3. AFRICA 6. SEACRIFOG
2.5	Infrastructure: Trust	A trusted repository infrastructure is desirable, addressing aspects of sustainability, governance, technology, and quality.	PG, SS, SI, GC	3. AFRICA
2.6	Policy: Flexibility	Flexibility is required in terms of policies and licenses to accommodate the wide variety of local legislation and maturity of participation in open science.	PG, SS	3. AFRICA
2.7	Policy: License Options	Open Licenses will have to be supplemented by a small number of standard licenses dealing with valid restrictions (ethics and privacy, commercial data, and government classified data).	PG, SS, SI	3. AFRICA
2.8	Governance: Sovereignty	Legislation or concerns in respect of data sovereignty may dictate the composition of the network and degree of federation of the data infrastructure	PG, SI, HN	3. AFRICA
2.9	Infrastructure: Energy	Energy costs are high, and availability is often poor. This has implications for backup strategies, power supply management, and for location and nature of data storage and processing infrastructure.	SI, HN	3. AFRICA
2.10	Infrastructure: Connectivity	Connectivity is also expensive with frequent interruptions. A survey of connectivity options in site locations will be required before decisions are made on network topology and data storage nodes.	SI, HN	3. AFRICA
2.11	Infrastructure: Human Capacity	Soft infrastructure (human resources, guidance, capacity building requirements) will be required and has implications for network topology and data storage nodes, cost of human resources and their efficiency.	SI, HN, GC	3. AFRICA
2.12	Infrastructure: Adaptation Focus	Africa needs adaptation rather than mitigation focus, and in addition to selection of variables and observation infrastructure, this also indicates	SI, HN, GC	3. AFRICA

		a need for decision and planning support tools.		
2.13	Governance: Mediation	In Africa, significantly more work will be required to communicate evidence-based decision and planning support to a large variety of communities, cultures and situations, and the need for community engagement is significant.	PG, GC	3. AFRICA

A.3 Global Initiatives: Considerations

These requirements are based on the concerns determined from Section 4.

#	Aspect	Design Consideration	Architecture	Source or Reference
3.1	Discoverability	Data and all associated metadata will be discoverable through catalogues and search engines, and data access and use conditions, including licenses, will be clearly indicated.	PG, SS	2: ENVRI 4: GEO, ILTER, WDS, USE
3.2	Accessibility	Data will be accessible via online services, including, at minimum, direct download but preferably user-customizable services for visualization and computation.	PG, SS	2: ENVRI 4: GEO, ILTER, WDS, USE
3.3	Usability: Schema	Data will be structured using encodings that are widely accepted in the target user community and aligned with organizational needs and observing methods, with preference given to non-proprietary international standards.	SS	4: GEO, ILTER, WDS
3.4	Usability: Re-Use	Data will be comprehensively documented, including all elements necessary to access, use, understand, and process, preferably via formal structured metadata based on international or community-approved standards. To the extent possible, data will also be described in peer-reviewed publications referenced in the metadata record. Datasets and services will be classified in respect of readiness (Publication, Analysis, Indicators) for specific re-use applications.	SS	4: GEO, ILTER, WDS
3.5	Usability: Provenance	Data will include provenance metadata indicating the origin and processing history of raw observations and derived products, to	SS	4: GEO, WDS

		ensure full traceability of the product chain.		
3.6	Usability: Quality	Data will be quality-controlled and the results of quality control shall be indicated in metadata; data made available in advance of quality control will be flagged in metadata as unchecked.	SS	2: ENVRI 4: GEO, WDS
3.7	Preservation: Loss Prevention	Data will be protected from loss and preserved for future use; preservation planning will be for the long term and include guidelines for loss prevention, retention schedules, and disposal or transfer procedures.	PG, SS	2: ENVRI 4: GEO, WDS
3.8	Preservation: Fixity	Data and associated metadata held in data management systems will be periodically verified to ensure integrity, authenticity and readability.	PG, SS	2: ENVRI 4: GEO, WDS
3.9	Curation: Active Curation	Data will be managed to perform corrections and updates in accordance with reviews, and to enable reprocessing as appropriate; where applicable this shall follow established and agreed procedures.	PG, SS	2: ENVRI 4: GEO, ILTER, WDS
3.10	Curation: Identifiers	Data will be assigned appropriate persistent, resolvable identifiers to enable documents to cite the data on which they are based and to enable data providers to receive acknowledgement of use of their data.	SS, SI	4: GEO, ILTER, WDS, USE
3.11	Curation: Citability/ License Monitoring	Citation metrics and support for proper citation via DOIs, and monitoring of license assent will be required	SS, SI, GC	4: ILTER, WDS, USE
3.12	Curation: Usage Metrics	Gathering information on searches, metadata and data downloads, use of visualisation and exploration tools, and links to VREs will be required.	SS, SI, GC	4: ILTER, USE
3.13	Curation: User Feedback	Users need to be able to rate metadata and data services, and provide structured and unstructured feedback.	SS, SI, GC	4: ILTER, USE
3.14	Curation: Identifiers	In addition to data PIDs, persistent identifiers are required for a variety of other elements of the research infrastructure: samples and specimens, researchers, institutions, instruments, sites and platforms. It is additionally required for other types of research outputs (code, protocols, algorithms, …) and for	SS, SI	4: ILTER, RDA, USE

		concepts (vocabularies, name services, …)		
3.15	Curation: Granularity	Guidance and agents are required to generate and maintain metadata, fixity, and citation support for dynamically updated datasets.	SS, SI, GC	4: ILTER, RDA, WDS
3.16	Infrastructure: Modularity and Service Orientation	There is significant support for a service-oriented architecture for global research data infrastructure - this allows a 'plug-and-play' based composition of applications and portals, utilising contributions from many	SS, SI, GC	4: ILTER, GEO, RDA
3.17	Interoperability: Syntactic	Agreed standards are needed for the service APIs to be used for metadata and data. Implement in infrastructure and guidance. Use compliant open source software projects.	SS, SI, GC	4: ILTER, GEO, WDS, USE
3.18	Interoperability: Schematic	The schema (XML, JSON, CSV, binary formats, …) to be supported for data and metadata exchange must be agreed and documented, implemented in infrastructure. Use compliant open source software projects.	SS, SI, GC	4: ILTER, GEO, WDS, USE
3.19	Interoperability: Semantic	Agree on sets of vocabularies and name services to use, and if absent, implement in SEACRIFOG vocabulary services.	PG, SS, SI, GC	4: ILTER, GEO, WDS, USE
3.20	Interoperability: Mediation	There is a growing need for mediation and brokering, due to two considerations ● Crosswalks between metadata/ data services in respect of syntax, semantics and schema ● Mediation of services that are not completely standards-compliant	SS, SI, GC	4: ILTER, GEO, RDA
3.21	Governance: Peer Review	Data infrastructure requires peer review periodically for certification as a trusted repository.	PG	4: WDS
3.22	Governance: Business Continuity	Agreements and arrangements needed to persist data holdings and its PID-based access if data infrastructure cannot continue operations - temporarily (failover and disaster recovery) and permanently (continued limited services)	PG, HN, SI, GC	4: WDS, USE
3.23	Governance: Sustainability	Adequate long-term funding - not project funding - for basic operations.	PG	4: WDS
3.24	Infrastructure: Open Source	Use open source software, projects, and formats whenever possible to limit future lack of access and maintenance	PG, SS, SI	4: WDS
3.25	Curation:	Workflow procedures must be documented and	PG, SS, SI,	4: WDS, USE

	Procedures	evidence that these are monitored must be available.	GC	
3.26	Digital Systems: Registration	Users need to be managed for a number of reasons and are assigned several roles. These range from identification of unique end users to record usage statistics, through more elaborate registration and authentication for contributors and administrators.	PG, SS, SI, GC	4: USE
3.27	Infrastructure: Observation Network Definition	SEACRIFOG will require a mechanism for recording observation infrastructure to datasets, similar to ILTER DEIMS.	PG, SS, SI, GC	4: USE, ILTER
3.28	Infrastructure: Modern Discovery	Infrastructure for discovery needs to be based on state-of-the-art implementations that can handle very large metadata collections, such as ElasticSearch or SOLR. In addition, indexed facets result in simpler implementation.	SS, SI, GC	4: USE
3.39	Curation: Workflows and Value Chains	Digital objects progress through a workflow (life cycle) that includes curation checklists as well as supporting value chain states (raw, publication-ready, analysis-ready, indicator-ready)	SS, SI, GC	2: ENVRI 4: USE, WDS
3.40	Infrastructure: Application	Allow visualisation, exploration, download, subsetting, processing, citation, and rating of datasets and services.	SS, SI, GC	4: USE

A.4 Existing Infrastructure: Considerations

These requirements are based on the concerns determined from Section 5. Includes information from ICOS, SAEON, and SASSCAL.

#	Aspect	Design Consideration	Architecture	Source or Reference
4.1	Infrastructure: Modularity	Create a portfolio of services for metadata and datasets, based on agreed standards,and develop lightweight, client-based software components and applications that can be reconfigured for multiple purposes.	SS, SI	5: SAEON
4.2	Infrastructure: Standard Harvesters	In a federated infrastructure, it is a requirement to automate the synchronisation of metadata, using standard interfaces.	SS, SI, GC	5: SAEON

4.3	Infrastructure: Harvestable	Expose metadata, in turn, to global infrastructures such as ILTER, GEO, Fluxnet, GBIF, and others	SS, SI, GC	5: SAEON
4.4	Infrastructure: Crowdsourcing, Citizen Science, and Event Scraping	There is limited scope for this in SEACRIFOG based on the defined essentiaessentiaessential variables, hence this may be regarded as a secondary requirement. Use open source and standard platforms where possible.	SS, SI, GC	5: SAEON
4.5	Infrastructure: Architecture	Federated, distributed architecture is available, with design and requirements documentation. All new release code is available in GitHub for re-use by e.g. SEACRIFOG	SS, SI, GC	5: SAEON
4.5	Infrastructure: Hosting	SAEON has distributed hardware, networking and security infrastructure in place, with daily backups (this can be adjusted if need be). Infrastructure is virtualised and can be made available for SEACRIFOG base infrastructure.	HN	5: SAEON
4.6	Infrastructure: SASSCAL	SASSCAL Data Portal components may be re-usable by SEACRIFOG. The SASSCAL Weathernet software,used to manage and disseminate meteorological observation data, may also be available. To be confirmed prior to finalisation of Deliverable 5.4.	SS, SI, GC	5: SASSCAL
4.7	Infrastructure: ICOS	ICOS has confirmed the availability of their Carbon Flux processing software stack for SEACRIFOG implementation, with a test platform planned for deployment on a SAEON virtual machine. This can serve as a basis for infrastructure of last resort in Africa for such measurements, and as a transportable installation image for similar deployments in Africa.	SS, SI, GC	5: ICOS

A.5 Considerations Derived from SEACRIFOG Observation Design

These requirements are based on the concerns determined from Section 6, and are based on the work done to collate the status of current observation across the essential variables in Africa.

#	Aspect	Design Consideration	Architecture	Source or Reference

5.1	Infrastructure: System Requirement	The SEACRIFOG Collaborative Inventory tool can be used as a basis for development of an observation inventory system similar to DEIMS.	SS, SI, HN, GC	6: SEACRIFOG
5.2	Infrastructure: Architecture	Such as system should link data and metadata that are curated elsewhere (for example instrumentation and RS mission metadata) based on a Linked Open Data model	SS, SI, GC	6: SEACRIFOG
5.3	Infrastructure: System Requirement	Contributors (networks, research institutions) should be able to administer and maintain the definition of their own observation infrastructure. If SEACRIFOG becomes a network-like institution, such maintenance may be a condition of membership,and hence completeness and update frequency must be monitored and auditable,	PG, SS, SI, GC	6: SEACRIFOG
5.4	Infrastructure: User Experience	End users should be able to explore available observation infrastructure and data on the basis of variable, instrument, location, temporal coverage, missions and initiatives, institutions, protocol, forcing, domain, etc., and be able to link via standardised services to distributed data associated with the infrastructure metadata.	SS, SI, GC	6: SEACRIFOG
5.5	Infrastructure: Big Data	A large proportion of the data implied for essential variables will be 'Big Data' - and this is best provisioned in cloud services, especially for remotely sensed data cubes and data requiring intensive pre-publication processing (carbon flux data)	SS, SI, GC, HN	6: SEACRIFOG
5.6	Infrastructure: Marine Data	Sovereignty issues are less pertinent and can be provisioned centrally.	SS, SI, GC, HN	6: SEACRIFOG
5.7	Infrastructure: High velocity in situ	For in situ observations with high velocity, provide centralised facilities in cases where it is not practical to host data infrastructure close to source, and invest in connectivity to enable transfer of high velocity data to a shared facility for processing and hosting	SS, SI, GC, HN	6: SEACRIFOG
5.8	Infrastructure: Non - 'Big Data'	Any other observations, provide centralised 'infrastructure of last resort' - but allow federated data hosting with centralised metadata aggregation and management.	SS, SI, GC, HN	6: SEACRIFOG

Annexure B. Data Families and Analysis of Variables

B.1 Data Families

Based on Mirtl et al. (2018)

Data Family	Typical Dimensionality	Typical Metadata Standards	Typical Data Services	Typical Operational Environment	Crosswalks
Traditional Spatial Data ("*Spatial*")	t, XYz, P, B, C	ISO 19115 FGDC	OGC WMS, WCS, WFS	Spatial Database, File System	Virtual WCS
Multidimensional Data ("*Cube*")	T, XYZ, P, B, C	ISO 19115	OpenDAP ErDDAP	NetCDF Array Database	WMS WCS Virtual WCS
Physico-Chemical Observations Data ("*Time Series*")	T, xyz, P, B, (C)	ISO 19115 SensorML	SensorThings SOS/ O&M	RDBMS Text Files NoSQL Databases	WxS Virtual WCS
Subtypes	Point: xyz Profile: xyz with one variable dimension Transect: xyz varying along a trajectory Coverage: xyz near-continuous (e.g. a raster)				
Ecosystem Observation Data [11] ("*Structured*")	T, xyz, (P), B, (C), Tx	EML DwC +	DwC + Object Download	MetaCAT RDBMS Spreadsheets Text Files Images Video Audio	Virtual WCS
Subtypes	Point: xyz Profile: xyz with one variable dimension Transect: xyz varying along a trajectory Coverage: xyz near-continuous (e.g. a raster)				
Genetic Data ("*Genetic*")	t, xyz, Al		FTP ASN.1	GenBank	Virtual WCS

The abbreviations in the table are:

- S-DB: spatial database;
- WxS: OGC (Open Geospatial Consortium web services);
- O&M: OGC Observations and Measurements model;
- SOS: OGC Sensor Observation Service;
- CSV: comma separated value;

[11] SAEON is exploring the concept of a 'BioCube' to serve as a virtual repository for all EBVs.

- DwC: Darwin Core, and DwC+ - with extensions,
- RDBMS - relational databases.,
- Virtual WCS - the ability to query and subset data in any data family as if it were an array database ("Data Cube").

The different types of coverage (spatial, temporal and semantic) and their attributes are:

- Spatial Coverage: XYZ (continuous), xyz (discrete), (xyz) (incidental)
- Temporal Coverage: T (continuous or near-continuous); t (discrete)
- Topic or Semantic/Ontological Coverage
    - Physico-Chemical Phenomenon: P (primary), (P) (incidental)
    - mostly physical, chemical, or other contextual data
    - Biological/ Ecosystem: B (primary), (B) (incidental)
    - aspects such as traits, biomass, occurrence, abundance, structure (EBVs)
    - Species and Taxonomy (with some extensions): Tx
    - Allele/Genome/Phylogenetic: Al
- The dimension of a sample, sampling event or specimen applies to all data families: S.

B.2 Variables and Assessment

Table: Essential variables as identified by SEACRIFOG and respective requirements for observation and data products. Note that uncertainties to ± 1 standard deviation from the actual value in percent, i.e. the percentual margins of a 68% confidence interval.

ID	Variable	Variable Class	Domain	Main Observation Technique	Observation Frequency	Spatial Res.	Max. Uncert.	Defined By	Data Family	Spatial Res. Class	Temporal Res. Class	Big Data
4	Area of ploughed land	Agricultural management	Terrestrial	Inventory/Census	5 years (resolve seasons)	100m	20%	SEACRIFOG	Spatial	High	Low	No
6	Irrigation	Agricultural management	Terrestrial	Combin. IS & RS	1 day	100 m	10%	SEACRIFOG	Spatial/ Cube	High	Medium	Yes
24	Burnt Area	Fire	Terrestrial	Remote sensing	1 day	30 m	15%	GCOS	Cube	High	Medium	Yes
32	Extent of inland waters	Land Cover	Terrestrial	Remote sensing	3 months	20 m	1%	SEACRIFOG	Cube	High	Medium	Yes
16	Tropospheric CH4 mixing ratio	Carbon Dioxide, Methane and other Greenhouse gases	Atmospheric	Combin. IS & RS	1 h	1 site	0.05% (1 ppb)	WMO GAW	Time Series	Low	High	Yes
17	Tropospheric CO2 mixing ratio	Carbon Dioxide, Methane and other Greenhouse gases	Atmospheric	Combin. IS & RS	1 h	1 site	0.25% (0.1 ppm)	WMO GAW	Time Series	Low	High	Yes
18	Tropospheric N2O mixing ratio	Carbon Dioxide, Methane and other Greenhouse gases	Atmospheric	Combin. IS & RS	1 h	1 site	0.05% (0.1 ppb)	WMO GAW	Time Series	Low	High	Yes

43	Non-methane hydrocarbons	Precursors	Atmospheric	In situ	1 h	1 site	10%	WMO GAW	Time Series	Low	High	Yes
40	Tropospheric Carbon Monoxide (CO)	Precursors	Atmospheric	Combin. IS & RS	1 h	1 site	1% (1 ppb)	WMO GAW	Time Series	Low	High	Yes
45	Pressure (surface)	Pressure (surface)	Terrestrial	In situ	1 h	1 Site	0.01% (0.1 hPa)	GCOS	Time Series	Low	High	Yes
57	Temperature (surface)	Temperature	Terrestrial	Combin. IS & RS	1 h	1 site	0.03% (0.1K)	GCOS	Time Series	Low	High	Yes
58	Water Vapour (surface)	Water Vapour (surface)	Terrestrial	In situ	1 h	1 site	1%	GCOS	Time Series	Low	High	Yes
11	Biosphere-Atmosphere CH4 flux	Biosphere-Atmosphere GHG flux	Terrestrial	In situ	1 h	1 site (every major ecoregion)	5%	SEACRIFOG	Time Series	Low	High	Yes
12	Biosphere-Atmosphere CO2 flux (NEE)	Biosphere-Atmosphere GHG flux	Terrestrial	In situ	1 h	1 site (every major ecoregion)	5%	SEACRIFOG	Time Series	Low	High	Yes
13	Biosphere-Atmosphere N2O flux	Biosphere-Atmosphere GHG flux	Terrestrial	In situ	1 h	1 site (every major ecoregion)	5%	SEACRIFOG	Time Series	Low	High	Yes
14	Boundary layer height	Boundary layer height	Atmospheric	Remote sensing	1 h	20 km	20%	SEACRIFOG	Time Series Cube	Low	High	Yes
19	Cloud Cover Fraction	Cloud Properties	Atmospheric	Remote sensing	1 h	25 km	10%	SEACRIFOG	Time Series Cube	Low	High	Yes
55	Surface Wind Speed and direction	Surface Wind	Terrestrial	Combin. IS & RS	3 h	10 km	10% (5 m/s)	GCOS	Time Series Cube	Low	High	Yes
3	Aerosol properties	Aerosol properties	Atmospheric	Combin. IS & RS	4 h	5 km	10%	GCOS	Time Series Cube	Low	High	Yes
42	Nitrogen Oxides (NOx)	Precursors	Atmospheric	Remote sensing	4 h	5 km	20%	GCOS	Time Series Cube	Low	High	Yes
44	Sulfur Dioxide (SO2)	Precursors	Atmospheric	Combin. IS & RS	4 h	5 km	30%	GCOS	Time Series Cube	Low	High	Yes
41	Oceanic Dimethyl Sulfide (DMS)	Precursors	Marine	In situ	1 month	1000 km	10%	SEACRIFOG	Time Series Cube	Low	Low	No
48	Net Radiation at surface (SW/LW)	Radiation	Terrestrial	Combin. IS & RS	1 month	100 km	0.25% (1 W/m2)	GCOS	Spatial/ Cube	Low	Low	No
38	Marine Oxygen	Marine Oxygen	Marine	In situ	1 month	100 km (marine	10%	GOOS	Spatial/ Cube	Low	Low	No

						biochemical province)						
31	Inorganic Carbon (Ocean)	Inorganic Carbon (Ocean)	Marine	In situ	1 month	250 km	10%	GOOS	Spatial/ Cube	Low	Low	No
20	Crop Yield by Type	Crops	Terrestrial	Inventory/Census	1 year	1 country	10%	SEACRIFOG	Spatial/ Cube	Low	Low	No
21	Economic Development	Economic Development	Marine	Inventory/Census	1 year	1 country	5%	SEACRIFOG	Spatial/ Cube	Low	Low	No
7	Manure Management	Agricultural management	Terrestrial	Inventory/Census	5 years	1 livestock system	20%	SEACRIFOG	Spatial	Low	Low	No
8	Livestock Distribution	Animal Population	Terrestrial	Inventory/Census	5 years	20 km	15%	SEACRIFOG	Spatial	Low	Low	No
9	Wild Herbivore Distribution	Animal Population	Terrestrial	Inventory/Census	5 years	20 km	15%	SEACRIFOG	Spatial	Low	Low	No
26	Human Population	Human Population	Terrestrial	Inventory/Census	5 years	20 km	5%	SEACRIFOG	Spatial	Low	Low	No
39	Plant Species Traits	Plant Species Traits	Terrestrial	In situ	Once off for all common species	All major biomes	10% (determine for 90% of cover)	SEACRIFOG	Structured	Low	Low	No
30	River Discharge	Hydrology	Terrestrial	In situ	1 day	1 river basin	10%	GCOS	TIme Series Cube	Low	Medium	No
15	Halocarbons	Carbon Dioxide, Methane and other Greenhouse gases	Atmospheric	In situ	1 week (flask)	1 site	5%	SEACRIFOG	Time Series	Low	Medium	No
5	Fertilizer application	Agricultural management	Terrestrial	Inventory/Census	1 year	1 country	20%	SEACRIFOG	Spatial	Low	Medium	No
53	Stable Carbon Isotopes	Stable Carbon Isotopes	Marine	In situ	3 months (seasonal)	100 km	10%	GOOS	Spatial/ Cube	Low	Medium	No
36	Marine Nutrients	Nutrients	Marine	In situ	3 months (seasonal)	100 km (marine biochemical province)	20%	GOOS	Spatial/ Cube	Low	Medium	No
23	Active Fire	Fire	Terrestrial	Remote sensing	1 h	250 m	5%	GCOS	Cube	Medium	High	Yes
22	Net Primary Productivity	Ecosystem Function	Terrestrial	Combin. IS & RS	1 month	1 km	10%	SEACRIFOG	Spatial/ Cube	Medium	Low	No
28	Infiltration and Runoff	Hydrology	Terrestrial	In situ	1 month	1 km	10%	SEACRIFOG	Spatial/ Cube	Medium	Low	No
46	Albedo	Radiation	Terrestrial	Remote sensing	1 month	300 m	5%	GCOS / SEACRIFOG	Spatial/ Cube	Medium	Low	No
25	Fire Fuel Load	Fire	Terrestrial	In situ	1 year	1 km	15%	SEACRIFOG	Spatial/ Cube	Medium	Low	No

34	Land Use/Land Use Change	Land Use/Land Use Change	Terrestrial	Combin. IS & RS	1 year	1 km	20%	SEACRIFOG	Spatial/ Cube	Medium	Low	No
33	Land Cover	Land Cover	Terrestrial	Remote sensing	1 year	250 m	15%	GCOS	Spatial/ Cube	Medium	Low	No
1	Above ground biomass	Above ground biomass	Terrestrial	Combin. IS & RS	1 year	500 m	20%	GCOS	Spatial/ Cube	Medium	Low	No
52	Soil Organic Carbon	Soil Properties	Terrestrial	In situ	10 years	1 km	5%	SEACRIFOG	Spatial	Medium	Low	No
10	Below-Ground Biomass	Below-Ground Biomass	Terrestrial	In situ	5 years	1 km	10%	SEACRIFOG	Spatial	Medium	Low	No
54	Surface Roughness	Surface Roughness	Terrestrial	In situ	5 years	1 km	20%	SEACRIFOG	Spatial	Medium	Low	No
2	Litter	Above ground biomass	Terrestrial	In situ	5 years (resolve seasons)	1 km	10%	SEACRIFOG	Spatial	Medium	Low	No
27	Evapotranspiration	Hydrology	Terrestrial	In situ	1 day	1 km	10%	SEACRIFOG	Spatial/ Cube	Medium	Medium	Yes
29	Precipitation (surface)	Hydrology	Terrestrial	Combin. IS & RS	1 day	1 km	10%	SEACRIFOG	Spatial/ Cube	Medium	Medium	Yes
35	Nitrous Oxide (Ocean)	Nitrous Oxide	Marine	In situ	1 day	1 km	1%	GOOS	Spatial/ Cube	Medium	Medium	Yes
51	Soil Moisture	Soil Properties	Terrestrial	Combin. IS & RS	1 day	1 km	4% (0.04 m3/m3)	GCOS	Spatial/ Cube	Medium	Medium	Yes
56	Sea Surface Temperature	Temperature	Marine	Combin. IS & RS	1 day	1 km	0.03% (0.1 K)	SEACRIFOG	Spatial/ Cube	Medium	Medium	Yes
37	Ocean Colour	Ocean Colour	Marine	Remote sensing	8 days	1 km	5%	SEACRIFOG	Spatial/ Cube	Medium	Medium	Yes
50	Sea Surface Salinity	Sea Surface Salinity	Marine	Combin. IS & RS	8 days	1 km	1%	SEACRIFOG	Spatial/ Cube	Medium	Medium	Yes
47	Fraction of Absorbed Photosynthetically Active Radiation (FAPAR)	Radiation	Terrestrial	Remote sensing	8 days	300 m	5%	SEACRIFOG	Spatial/ Cube	Medium	Medium	Yes
49	CO2, CH4, N2O emissions by country and IPCC sector	Reported Anthropogenic Greenhouse Gas Emissions	Terrestrial	Inventory/ Census	1 year	1 country	10%	SEACRIFOG	Spatial/ Cube	Low	Low	No